

PLEASE TYPE**THE UNIVERSITY OF NEW SOUTH WALES**
Thesis/Dissertation SheetSurname or Family name: **Yau**First name: **Sheree**

Other name/s:

Abbreviation for degree as given in the University calendar: **PhD**School: **Biotechnology and Biomolecular Sciences**Faculty: **Faculty of Science**Title: **Molecular studies of saline Antarctic lakes from a whole ecosystem perspective****Abstract 350 words maximum: (PLEASE TYPE)**

The Vestfold Hills is a coastal Antarctic oasis, a rare ice-free region on the continent containing hundreds of marine-derived lakes. These lakes are microbially-dominated systems constrained by extremes of cold, salinity and light availability. Differing local geographic features has led each lake to develop unique chemistries tightly linked to the resident microbial populations. Most Antarctic lakes are ice-covered for a large proportion of the year and so are largely closed systems. Their physical isolation and the extreme environmental constraints imposed by the polar environment makes them potential reservoirs of novel taxa. Furthermore, many lakes are stratified so the physical and chemical gradients that exist within a single system makes it possible to relate microbial taxa to physico-chemical variables. These factors make Antarctic lakes ideal model ecosystems in which to study microbial diversity, evolution and influence on geochemistry.

It is now well-known most microbes are not readily culturable by standard methods. Application of molecular biology techniques to the study of microbes directly from the environment has revealed the enormous diversity of microbial life. In particular, random high-throughput sequencing of genetic material directly from the environment (metagenomics) allows the determination not only of the microbial community composition and structure, but also their metabolic potential. The studies described in this thesis use metagenomic sequencing of microbial communities in stratified saline lakes in the Vestfold Hills. In order to describe the role of microbes in the ecology of the whole lake environment, analyses complementary to metagenomic sequencing were also developed and applied. These included identification and quantification of proteins extracted directly from the environment (metaproteomics) and microscopy for direct counts of microbial and viral abundances and morphological examination.

Taken in combination with both historic and contemporary physico-chemical data, molecular-based information allowed a description of the lake ecosystem and also resulted in totally new insights into mechanisms of adaption to the Antarctic environment. From these genomic discoveries, hypotheses of the role previously unknown taxa and functional genes have on the environment were developed and modelled. These discoveries not only have implications for Antarctic environments, but also for other aquatic systems.

Declaration relating to disposition of project thesis/dissertation

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).

.....
Signature.....
Witness.....
Date

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

FOR OFFICE USE ONLY

Date of completion of requirements for Award:

THIS SHEET IS TO BE GLUED TO THE INSIDE FRONT COVER OF THE THESIS

Molecular microbial ecology of Antarctic lakes

Sheree Yau

A thesis in fulfilment of the requirements for the degree of Doctor of Philosophy

School of Biotechnology and Biomolecular Sciences
Faculty of Science
University of New South Wales, Australia

January, 2013

Originality Statement

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed

Date

Authenticity Statement

'I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.'

Acknowledgements

List of Publications

Publications and submitted manuscripts arising from my PhD research are listed below. In all cases, my supervisor Prof Ricardo Cavicchioli and my co-supervisor Dr Federico Lauro were involved in the research design and editing of the manuscripts. Where versions of published material, or material submitted for publication appears in this thesis, details of the contributions made by myself and others precede it.

- **Sheree Yau**, Federico M. Lauro, Timothy J. Williams, Matthew Z. DeMaere, Mark V. Brown, John Rich, John A.E. Gibson, Ricardo Cavicchioli. Strategies of carbon conservation and unusual sulfur biogeochemistry in a hypersaline lake. *ISME Journal* (submitted), 2013.
- Khawar S. Siddiqui, Timothy J. Williams, David Wilkins, **Sheree Yau**, Michelle A. Allen, Mark V. Brown, Federico M. Lauro, Ricardo Cavicchioli. Psychrophiles. *Annual Review of Earth and Planetary Sciences* (in press), 2013.
- David Wilkins, **Sheree Yau**, Timothy J. Williams, Michelle Allen, Mark V. Brown, Matthew Z. DeMaere, Federico M. Lauro and Ricardo Cavicchioli. Key Microbial Drivers in Antarctic Aquatic Environments. *FEMS Microbiology Reviews* (10.1111/1574-6976.12007), 2012.
- **Sheree Yau** and Ricardo Cavicchioli. Microbial communities in Antarctic lakes: Entirely new perspectives from metagenomics and metaproteomics. *Microbiology Australia* 32:157–159, 2011.
- Federico M. Lauro, Matthew Z. DeMaere, **Sheree Yau**, Mark V. Brown, Charmaine Ng, David Wilkins, Mark J. Raftery, John A.E. Gibson, Cynthia Andrews-Pfannkoch, Matthew Lewis, Jeffery M. Hoffman, Torsten Thomas and Ricardo Cavicchioli. An integrative study of a meromictic lake ecosystem in Antarctica. *ISME Journal* 5:879–895, 2011.
- **Sheree Yau**, Federico M. Lauro, Matthew Z. DeMaere, Mark V. Brown, Torsten Thomas, Mark J. Raftery, Cynthia Andrews-Pfannkoch, Matthew Lewis, Jeffery M. Hoffman, John A. Gibson and Ricardo Cavicchioli. Virophage control of antarctic algal host–virus dynamics. *Proceedings of the National Academy of Sciences USA* 108:6163–6168, 2011.

Contents

1 General introduction	1
Co-authorship statement	1
1.1 Antarctic lakes	2
1.2 The Vestfold Hills, East Antarctica	3
1.3 Molecular approaches used in Antarctic lake systems	3
1.4 Insights from Antarctic molecular studies	5
1.4.1 Salinity affects bacterial diversity	5
1.4.2 <i>Archaea</i> : methanogens and haloarchaea	7
1.4.3 <i>Eucarya</i> perform multiple ecosystem roles	8
1.5 Limitations of taxonomic surveys	9
1.5.1 Functional gene studies of Antarctic lakes	9
1.5.2 Integrative studies to derive whole ecosystem function	10
1.6 ‘-omics’ approaches	10
1.7 Objectives	10
2 Development of methods to complement metagenomic analysis of Ace Lake	13
Co-authorship statement	13
2.1 Abstract	15
2.2 Introduction	15
2.3 Materials and methods	16
2.3.1 Ace Lake samples	16
2.3.2 DNA sequencing and data cleanup	17
2.3.3 Metagenomic DNA assembly and annotation	18
2.3.4 Epifluorescence microscopy	18
2.3.5 Metaproteomic analysis	18
2.4 Results and discussion	20
2.4.1 Epifluorescence microscopy methodology	20
2.4.2 Community stratification supported by cell and virus-like particle (VLP) densities	23
2.4.3 Development of metaproteomic methodology	24
2.5 Conclusions	25

3 Virophage control of Antarctic algal host–virus dynamics	27
Co-authorship statement	27
3.1 Abstract	28
3.2 Introduction	29
3.3 Materials and methods	30
3.3.1 Samples and DNA sequencing	30
3.3.2 Transmission electron microscopy	30
3.3.3 Metagenomic assembly and annotation	30
3.3.4 Genome completion and annotation	31
3.3.5 Phylogenetic analysis	32
3.3.6 Metaproteomic analysis	32
3.3.7 Algal Host–Virus and Virophage Dynamics	32
3.4 Results and discussion	33
3.4.1 Dominance of phycodnaviruses in Organic Lake	33
3.4.2 Complete genome of an Organic Lake virophage	38
3.4.3 Gene exchange between virophage and phycodnaviruses	45
3.4.4 Virophage in algal host–phycodnavirus dynamics	46
3.4.5 Ecological relevance of virophages in aquatic systems	48
3.5 Acknowledgements	51
4 Strategies of carbon conservation and unusual sulphur biogeochemistry in a hypersaline Antarctic Lake	53
4.1 Abstract	54
4.2 Introduction	55
4.3 Materials and methods	56
4.3.1 Characteristics of the lake and sample collection	56
4.3.2 Physical and chemical analyses	56
4.3.3 Epifluorescence microscopy	57
4.3.4 Cellular diversity analyses	58
4.3.5 Analysis of functional potential	58
4.4 Results and discussion	62
4.4.1 Abiotic properties and water column structure	62
4.4.2 Overall microbial diversity	67
4.4.3 Variation of microbial composition according to size and depth	67
4.4.4 Organic Lake functional potential	73
4.4.5 Carbon resourcefulness in dominant heterotrophic bacteria	73
4.4.6 Regenerated nitrogen is predominant in the nitrogen cycle	80
4.4.7 Molecular basis for unusual sulphur chemistry	83
4.4.8 Conclusion	93
4.5 Acknowledgements	94
5 General discussion, conclusions and future work	95

List of Figures

1.1	Map of the Vestfold Hills	4
2.1	Epifluorescence microscopy images of Ace Lake microbiota	22
2.2	Counts of microbial cells and VLPs by epifluorescence microscopy	23
3.1	Plot of percent GC content vs coverage for the 2006 Organic Lake 0.1 µm hybrid assembly scaffolds >10 kbp	34
3.2	Transmission electron micrographs of VLPs	35
3.3	Phylogeny of Organic Lake phycodnavirus (OLPV) B family DNA polymerase sequences	36
3.4	Phylogeny of the 18S rRNA genes from the 2006 Organic Lake metagenome	37
3.5	Maps of OLPV genomic scaffolds	38
3.6	Phylogeny of OLPV major capsid sequences	39
3.7	Genomic map of Organic Lake virophage	41
3.8	Comparison of the location of genes in Organic Lake virophage (OLV) compared to OLPVs	46
3.9	Extended Lotka-Volterra models of host–OLPV–OLV population dynamics	47
3.10	Abundance and diversity of virophage capsid proteins in environmental samples	48
3.11	Phylogeny of virophage capsid proteins	50
4.1	Vertical profiles of <i>in situ</i> Organic Lake abiotic parameters	62
4.2	Bathymetry of Organic Lake	63
4.3	Vertical structure of Organic Lake	64
4.4	Epifluorescence microscopy images of Organic Lake microbiota	65
4.5	PCA analysis of physico-chemical parameters	66
4.6	Diversity of <i>Bacteria</i> in Organic Lake	68
4.7	Diversity of <i>Eucarya</i> in Organic Lake	69
4.8	Heatmap and biclustering plot of the small subunit ribosomal RNA (SSU) gene composition in Organic Lake	70
4.9	Vertical profiles of potential for carbon cycling in Organic Lake	75
4.10	Phylogenetic tree of rhodopsin homologues	77
4.11	Genomic maps of Organic Lake scaffolds containing the OL-R1 rhodopsin homologue	78
4.12	Vertical profiles of potential for nitrogen cycling in Organic Lake	81

4.13	Vertical profiles of potential for sulphur cycling in Organic Lake	84
4.14	Phylogenetic tree of DddD DMSP lyase homologues	87
4.15	Genomic maps of Organic Lake scaffolds containing the OL-dddD homologue	88
4.16	Phylogenetic tree of DddL DMSP lyase homologues	89
4.17	Phylogenetic tree of DddP DMSP lyase homologues	90
4.18	Phylogenetic tree of DmdA DMSP demethylase homologues	91

List of Tables

3.1	Summary of metagenomic data for Organic Lake 0.1 μm samples used in this study	30
3.2	List of primers used to close the OLV genome.	31
3.3	OLPV and OLV proteins identified in the metaproteome	40
3.4	Top BLASTP matches of predicted coding sequences from the Organic Lake virophage genome	42
3.5	BLASTP matches for OLV major capsid protein (MCP) in predicted open-reading frames (ORFs) of Organic Lake and Ace Lake contigs and Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) metagenomic read ORF peptide database	49
4.1	Summary of metagenomic data for Organic Lake profile	57
4.2	List of KEGG orthologs (KO) involved in nutrient cycles that were searched for in the Organic lake metagenome	59
4.3	List of query functional marker gene sequences	61
4.4	Physico-chemical properties of Organic Lake 2008 vertical profile	66
4.5	Contribution of different taxonomic groups to counts of marker genes involved in carbon conversions	74
4.6	Counts of genes involved in dimethylsulphopropionate (DMSP) catabolism and photoheterotrophy in aquatic metagenomes	79
4.7	Contribution of different taxonomic groups to counts of marker genes involved in nitrogen conversions	82
4.8	Contribution of different taxonomic groups to counts of marker genes involved in sulphur conversions	86
6.1	Peptide data for Organic Lake metaproteomics analysis	109
6.2	Proteins identitifed in the Ace Lake 5 m sample 0.1 μm size-fraction proteome	111

List of Abbreviations

1D-SDS PAGE one dimensional-sodium dodecyl sulphate polyacrylamide gel electrophoresis

AAnP aerobic anoxygenic photosynthesis

ABC ATP-binding cassette

ACT Artemis comparison tool

ANOSIM analysis of similarity

ApMV *Acanthamoeba polyphaga* mimivirus

BchlA bacteriochlorophyll A

BLAST basic local alignment search tool

CAMERA Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis

CAS CRISPR-associated proteins

COG clusters of orthologous groups

CRISPR clustered regularly interspaced short palindromic repeat

DGGE denaturing gradient gel electrophoresis

DMS dimethylsulphide

DMSO dimethylsulphoxide

DMSP dimethylsulphopropionate

DPOB DNA polymerase B

DO dissolved oxygen

DOC dissolved organic carbon

DRP dissolved reactive phosphorus

GOS global ocean sampling

- GSB** green sulphur bacteria
- HMMER** biosequence analysis using profile hidden Markov models
- JCVI** J.Craig Venter Institute
- KEGG** Kyoto Encyclopedia of Genes and Genomes
- KO** KEGG Orthology
- KOBAS** KEGG Orthology Based Annotation System
- MCP** major capsid protein
- MEGA** Molecular Evolutionary Genetic Analysis
- MS** mass spectrometry
- MS-MS** two dimensional mass spectrometry
- NCBI** National Center for Biotechnology Information
- NCLDV** nucleo-cytoplasmic large DNA virus
- NR** non-redundant database
- OLPV** Organic Lake phycodnavirus
- OLV** Organic Lake virophage
- ORF** open-reading frame
- OTU** operational taxonomic unit
- PCA** principal component analysis
- PCR** polymerase chain reaction
- PCTE** polycarbonate Track EtchTM
- PR** proteorhodopsin
- PV** phycodnavirus
- QIIME** Quanitative Insights Into Microbial Ecology
- RDP** Ribosomal Database Project
- RE** restriction endonuclease
- RM** restriction modification
- rRNA** ribosomal RNA
- rTCA** reverse tricarboxylic acid

RuBisCO ribulose-bisphosphate carboxylase oxygenase

SO Southern Ocean

SRB sulphate-reducing bacteria

SSU small subunit ribosomal RNA

STAMP Statistical Analysis of Metagenomic Profiles

TEM transmission electron microscopy

TDN total dissolved nitrogen

TDP total dissolved phosphorus

TDS total dissolved sulphur

TIGRFAM the Institute of Genomic Research curated protein database

TN total nitrogen

TOC total organic carbon

TP total phosphorus

TS total sulphur

VLP virus-like particle

WGS whole genome shotgun

WL Wood-Ljungdahl; or reductive acetyl-CoA

Chapter 1

General introduction: molecular microbial ecology of Antarctic lakes

Co-authorship statement

Sections of this chapter has been published as:

David Wilkins, **Sheree Yau**, Timothy Williams, Michelle Allen, Mark V. Brown, Matthew Z. DeMaere, Federico M. Lauro and Ricardo Cavicchioli. Key Microbial Drivers in Antarctic Aquatic Environments. *FEMS Microbiology Reviews* (doi: 10.1111/1574-6976.12007), 2012.

I contributed the section of the publication entitled *Antarctic lakes* excluding the subsection, *Microbial mats as microcosms of Antarctic life*. This material appears in section 1.3 *Molecular approaches used in Antarctic lake systems*, section 1.4 *Molecular insights into Antarctic lakes* and section 1.5 *limitations of taxonomic surveys* of this introduction.

Antarctica is a “frozen desert” of constant low temperature, little precipitation and subject to the polar light cycle where only specially adapted organisms can survive. The continent is covered by ice up to 4 km thick that spans 13.8 million km². A tiny 0.32% of the land area is ice-free, most of which consists of exposed rocky peaks or nunataks such as in the Ellsworth, the Transantarctic and the North Victoria Land Mountains. Only 1–2% of that ice-free land is found in coastal oases; however, it is these regions where Antarctic life is concentrated (Hodgson, 2012).

They are breeding sites for large animals such as seals, penguins and sea birds and some of the only locations where plants and lichens are found. Coastal oases are also distinguished by the presence of hundreds of lakes and ponds. Life in these lakes is microbially dominated with few, if any, metazoan inhabitants (Laybourn-Parry, 1997) making them ideal locations to study Antarctic microbiota. The lakes span a continuum of environmental factors such as salinity and are “natural laboratories” to examine adaptations to a property of interest. They are also ideal model ecosystems as they are largely isolated with a close relation between species and function.

This introduction will describe the Antarctic lakes, their microbiology and review molecular-based Antarctic microbiological research on the lakes. As this thesis focused on two lakes in the Vestfold Hills, emphasis will be given to describing research from this study site.

1.1 Antarctic lakes

In Antarctica, perennially available liquid water is found predominantly in lakes. The majority of lakes are found in the coastal oases. In East Antarctica these include the Vestfold Hills, Bunker Hills, Larsemann Hills, Syowa Oasis, Schirmacher Oasis, Grearson Hills and McMurdo Dry Valleys. In West Antarctic, the Antarctic Peninsula, the subantarctic islands and maritime islands house multiple lakes. Of these locations, the best studied lake systems are those of the McMurdo Dry Valleys, The Vestfold Hills and the subantarctic islands.

These lakes span a wide range of physical and chemical properties from freshwater to hypersaline and constantly ice-covered to melted. Some are permanently stratified and termed meromictic if they thaw seasonally, or amictic if they are always ice-covered. Stratified lakes provide a unique opportunity to describe microbial populations along chemical gradients within a single water body. Most lakes are ice-covered for most of the year making them effectively isolated, and some may be truly closed systems if ice-cover is permanent. The age of water varies considerably; for example, outflow of subglacial water at Blood Falls is estimated to be 1.5 million years old (Mikucki *et al.*, 2009) while water from Lake Miers is less than 300 years old (Green *et al.*, 1988). Overall, there are two main Antarctic lake types: those bound by ice, comprising subglacial; epiglacial and supraglacial lakes, and those bound by rock.

Most of these lakes were formed when the retreat of the continental ice-shelf lead to isostatic uplift of the land (Burton, 1981). As a result, the majority of lakes in the coastal oases are composed of relic seawater and are predominantly saline or hypersaline

(Burke and Burton, 1988). In the latter, salinity is high due to concentrated by ablation (evaporation and sublimination). Lakes closer to the coastline may still occasionally experience marine inputs.

Freshwater lakes near the continental ice shelf were likely already above sea-level as the ice receded and are not of marine origin (Bronge, 2004). Other freshwater lakes were originally marine-derived but have been flushed fresh by glacial meltwater (Pickard *et al.*, 1986). The chemistry of the exposed lakes is very much influenced by the water balance from local geographic and climatic conditions which leads them to have different physical and chemical properties. Input sources include precipitation, from the ice-shelf and glacial melt streams (Burton, 1981).

1.2 The Vestfold Hills, East Antarctica

The Vestfold Hills (Figure 1.1) is a ice-free region of approximately 400 km² on the eastern shore of the Prydz Bay, East Antarctica in the Australian Antarctic Territory (Gibson, 1999). The region is made up of three large peninsulae, Broad, Mule and Long Peninsula, separated by fjords connected to the sea. Some of these are large, such as Ellis Fjord, which is 10 km long, up to 100 m deep and has become a stratified system due to its restricted opening to the ocean (Burke and Burton, 1988). The region was formed approximately 10,000 years ago in the early Holocene as the continental ice receded and the rocky peninsulae rose above sea-level (Zwartz *et al.*, 1998).

The Vestfold Hills were first sighted and named in 1935 (Law, 1959). Only intermittent expeditions occurred in the area until the establishment of Davis Station (68°33'S, 78°15'E) in 1957 (Law, 1959). A continuous presence has been maintained since. The Vestfold Hills region was immediately noted for its extensive ice-free land and the numerous lakes (Johnstone *et al.*, 1973).

The Australian Antarctic Data Centre lists more than 3,000 water bodies mapped in the Vestfold Hills, ranging in area from 1 to 8,757,944 m². More than 300 lakes and ponds have been described, including approximately 20% of the world's meromictic lakes (Gibson, 1999).

The biota of the lakes is largely microbial. Metazoan life is absent from many lakes with copepods being the largest organisms present. Thus, Antarctic lakes are ideal sites to study microbial ecology.

These are of particular interest because the anoxic bottom waters help preserve a paleogeological record in the sediments. This can tell us about the region and particularly climatic changes. Stratified lakes provide a unique opportunity to describe microbial populations along chemical gradients, but within a single water body.

1.3 Molecular approaches used in Antarctic lake systems

The majority of molecular-based studies of Antarctic aquatic microbial communities have made use of polymerase chain reaction (PCR) amplification of SSU sequences to survey the diversity of *Bacteria* and in some cases *Archaea* and *Eucarya* (??). Microbial

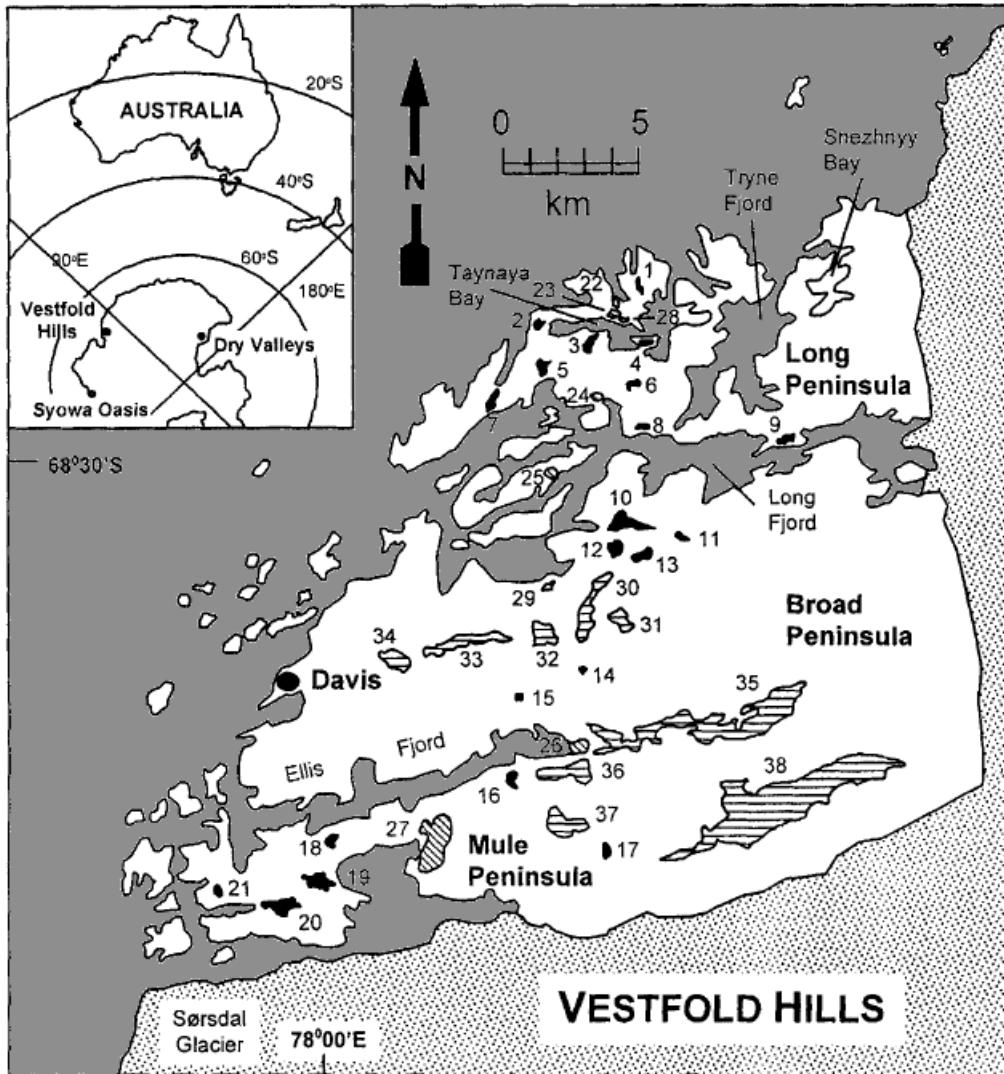


Figure 1.1: A map of the Vestfold Hills showing fjords, bays and lakes (numbered). The Southern Ocean is shown in grey, meromictic lakes coloured in black, seasonally isolated lakes and basins are striped and the continental ice-shelf is stippled. Inset is the position of the Vestfold Hills relative to Australia and to the Antarctic coastal oases. The lakes are: (1) unnamed lake 2, (2) Organic Lake, (3) Pendant Lake, (4) Glider Lake, (5) Ace Lake, (6) unnamed lake 1, (7) Williams Lake, (8) Abraxas Lake, (9) Johnstone Lake, (10) Ekho Lake, (11) Lake Farrell, (12) Shield Lake, (13) Oval Lake, (14) Ephyra Lake, (15) Scale Lake, (16) Lake Anderson, (17) Oblong Lake, (18) Lake McCallum, (19) Clear Lake, (20) Laternula Lake and (21) South Angle Lake. Map image modified from (Gibson, 1999).

composition has been determined by cloning and sequencing of ribosomal RNA (rRNA) gene amplicons exclusively (Bowman *et al.*, 2000b,a; Gordon *et al.*, 2000; Christner *et al.*, 2001; Purdy *et al.*, 2003; Karr *et al.*, 2006; Matsuzaki *et al.*, 2006; Kurosawa *et al.*, 2010; Bielewicz *et al.*, 2011), although most studies have also made use of denaturing gradient gel electrophoresis (DGGE) to provide a molecular “fingerprint” of the community (Pearce, 2003; Pearce *et al.*, 2003; Karr *et al.*, 2005; Pearce, 2005; Pearce *et al.*, 2005; Unrein *et al.*, 2005; Glatz *et al.*, 2006; Mikucki and Priscu, 2007; Mosier *et al.*, 2007; Schiaffino *et al.*, 2009; Villaescusa *et al.*, 2010). Functional genes have also been targeted using PCR amplification to assess the potential of biochemical processes occurring, such as nitrogen fixation (Olson *et al.*, 1998), ammonia oxidation (Voytek *et al.*, 1999), anoxygenic photosynthesis (Karr *et al.*, 2003), and dissimilatory sulfite reduction (Karr *et al.*, 2005; Mikucki *et al.*, 2009).

1.4 Insights from Antarctic molecular studies

The vast majority of molecular studies of Antarctic lakes have focused on *Bacteria*, although *Archaea*, *Eucarya* and viruses are also starting to be examined this way. Consistent with the wide range of physical and chemical properties of Antarctic lakes, a large variation in species assemblages have been found. While exchange of microorganisms must be able to occur between lakes that are in close vicinity to each other, the data to date indicates that microbial populations are relatively unique to each type of isolated system. Nonetheless, certain trends in bacterial composition are also apparent.

1.4.1 Salinity affects bacterial diversity

Lakes of equivalent salinities tend to have similar communities. Hypersaline lakes from the Vestfold Hills (Bowman *et al.*, 2000a) and McMurdo Dry Valleys (Glatz *et al.*, 2006; Mosier *et al.*, 2007) were all dominated by *Gammaproteobacteria* and members of the *Bacteroidetes* as well as harboring lower abundance populations of *Alphaproteobacteria*, *Actinobacteria*, and *Firmicutes*. The surface waters of lakes close to marine salinity resemble marine communities dominated by *Bacteroidetes*, *Alphaproteobacteria* and *Gammaproteobacteria*, but divisions such as *Actinobacteria* and specific clades of *Cyanobacteria* have been found to be overrepresented compared to the ocean (Lauro *et al.*, 2011). Sediments from saline lakes in the Vestfold Hills (Bowman *et al.*, 2000b) and Nuramake-Ike in the Syowa Oasis (Kurosawa *et al.*, 2010) were very similar, containing in addition to the surface clades, *Deltaproteobacteria*, *Planctomycetes*, *Spirochaetes*, *Chloroflexi* (green non-sulphur bacteria), *Verrucomicrobia* and representatives of candidate divisions. Plankton from freshwater lakes were characterized by an abundance of *Betaproteobacteria*, although *Actinobacteria*, *Bacteroidetes*, *Alphaproteobacteria* and *Cyanobacteria* were also prominent (Pearce, 2003, 2005; Pearce *et al.*, 2005; Schiaffino *et al.*, 2009).

Nutrients affect bacterial diversity

Differences in bacterial community structure are also influenced by nutrient availability. In studies of freshwater lakes in the Antarctic Peninsula and the South Shetland Islands, cluster analysis of DGGE profiles grouped together lakes of similar trophic status (Schiaffino *et al.*, 2009; Villaescusa *et al.*, 2010). Most of the variance in community structure could be explained by related chemical parameters such as phosphate and dissolved inorganic nitrogen. Similarly, three freshwater lakes, Moss, Sombre and Heywood on Signy Island are alike except that Heywood Lake is enriched by organic inputs from seals.

Bacterial composition in each lake changed from winter to summer and this was again correlated to variation in physico-chemical properties (Pearce, 2005). The bacterial population of Heywood Lake had shifted from a dominance of *Cyanobacteria* towards a greater abundance of *Actinobacteria* and marine *Alphaproteobacteria* (Pearce *et al.*, 2005). This hints at a link between a copiotrophic lifestyle in the Heywood Lake *Actinobacteria* and inhibition of Antarctic freshwater *Cyanobacteria* by eutrophication. This type of study exemplifies how inferences can be made about taxa and function by examining population changes over time and over gradients of environmental parameters.

Bacterial biogeography

The relative isolation and diverse chemistries of the lakes facilitates biogeographical and biogeochemical studies. The anoxic and sulfidic bottom waters of some meromictic lakes form due to a density gradient that precludes mixing. Although sedimentation from the upper aerobic waters may occur, there is little opportunity for interchange of species with the bottom water of lakes allowing for greater divergence in community composition as nutrients can become depleted and products of metabolism can accumulate. As a result, distinct distributions of bacterial groups can inhabit these strata, and different types of microorganisms can be found in equivalent strata in different lakes. A good example of this is the presence of common types of purple sulphur bacteria (*Chromatiales*) and green sulphur bacteria (GSB) (*Chlorobi*) in some meromictic lakes and stratified fjords in the Vestfold Hills (Burke and Burton, 1988), compared to diverse purple non-sulphur bacteria in Lake Fryxell in Victoria Land (Karr *et al.*, 2003). In Lake Bonney, the east and west lobes harbor overlapping but distinct communities in the suboxic waters (Glatz *et al.*, 2006). The east lobe was dominated by *Gammaproteobacteria* and the west lobe by *Bacteroidetes*, illustrating how divergent communities can form from the same seed population. In contrast, ice communities are more readily dispersed by wind, aerosols and melt-water. 16S rRNA gene probes designed from bacteria trapped in the permanent ice-cover of Lake Bonney hybridized to microbial mat libraries sourced up to 15 km away (Gordon *et al.*, 2000). This demonstrates how a single lake may encompass microorganisms that are geographically dispersed, while also harboring others that have restricted niches and are under stronger selection pressure.

Bacterial diversity of Lake Vostok

Subglacial systems, such as Lake Vostok, have been isolated from the open environment for hundreds of thousands to millions of years (Siegert *et al.*, 2001). As a result they provide a reservoir of microorganisms that may have undergone significant evolutionary divergence from the same seed populations that were not isolated by the Antarctic ice cover. Lake Vostok is approximately 4 km below the continental ice-sheet making it extremely difficult to determine suitable means for accessing the lake without inadvertently contaminating it with biological or chemical matter (?????). To date, molecular microbial studies have concentrated on the accretion ice above the ice-water interface (Priscu, 1999; Christner *et al.*, 2001). Accretion ice has been found to contain a low density of bacterial cells from *Alphaproteobacteria*, *Betaproteobacteria*, *Actinobacteria* and *Bacteroidetes* divisions closely allied to other cold environments. Molecular signatures of a thermophilic *Hydrogenophilus* species were also identified in accretion ice raising the possibility that chemoautotrophic thermophiles were delivered to the accretion ice from hydrothermal areas in the lakes bedrock (Bulat *et al.*, 2004; ?). However, interpretation of results from samples sourced from the Lake Vostok bore hole are very challenging as it is difficult to differentiate contaminants from native Vostok microorganisms. From a study that assessed possible contaminants present in hydrocarbon-based drilling fluid retrieved from the Vostok ice core bore hole, six phylotypes were designated as new contaminants (Alekhina *et al.*, 2007). Two of these were *Sphingomonas* phylotypes essentially identical to those found in the accretion ice-core (Christner *et al.*, 2001), which raises question about whether bacterial signatures identified from the ice-cores are representative of Lake Vostok water, and further highlights the ongoing problem of causing forward contamination into the lake.

1.4.2 Archaea: methanogens and haloarchaea

Archaea have been detected mainly in anoxic sediments and bottom waters from lakes that range in salinity from fresh to hypersaline, and those with known isolates are affiliated with methanogens or haloarchaea (Bowman *et al.*, 2000a,b; Purdy *et al.*, 2003; Kurosawa *et al.*, 2010; Lauro *et al.*, 2011). Anoxia allows for the growth of methanogenic *Archaea* that mineralize fermentation products such as acetate, H₂ and CO₂ into methane, thereby performing an important step in carbon cycling. The acetoclastic methanogens thrive in environments where alternative terminal electron acceptors such as sulfate and nitrate have been depleted.

One example of this is Lake Heywood where methanogenic *Archaea* were found to comprise 34% of the total microbial population in the freshwater sediment, the majority of which were *Methanosarcinales* which include acetate and C1-compound utilizing methanogens (Purdy *et al.*, 2003).

In general, archaeal populations appear to be adapted to their specific lake environment. Sediments from saline lakes of the Vestfold Hills were inhabited by members of the *Euryarchaeota* typically found in sediment and marine environments with the phylotypes differing between the lakes examined (Bowman *et al.*, 2000b). While a phylotype

similar to *Methanosarcina* was identified, the majority were highly divergent. Similarly, *Methanosarcina* and *Methanoculleus* were detected in Lake Fryxell but other members of the *Euryarchaeota* and *Crenarchaeota* (a single sequence) were divergent, clustering only with marine clones (Karr *et al.*, 2006). Based on the lake chemical gradients and the location of these novel phylotypes in the water column the authors speculated these archaea may have alternative metabolisms such as anoxic methanotrophy or sulphur-utilization.

In sediments from Lake Nurume-Ike in the Langhovde region, 205 archaeal clones grouped into three phylotypes, with the predominant archaeal clone being related to a clone from Burton Lake in the Vestfold Hills, while the other two did not match to any cultivated species (Kurosawa *et al.*, 2010). In hypersaline lakes where bottom waters do not become completely anoxic, methanogens are not present and *Archaea* have extremely low abundance. For example, only two archaeal clones of the same phylotype were recovered from deep water samples from Lake Bonney (Glatz *et al.*, 2006), and Organic Lake in the Vestfold Hills had an extremely low abundance of archaeal clones related to *Halobacteriales* (Bowman *et al.*, 2000a). In contrast to these stratified hypersaline lakes, the microbial community in the extremely hypersaline Deep Lake is dominated by haloarchaea (Bowman *et al.*, 2000a). Many of the clones identified from Deep Lake are similar to *Halorubrum* (formerly *Halobacterium*) *lacusprofundi* which was isolated from the lake (Franzmann *et al.*, 1988).

1.4.3 *Eucarya* perform multiple ecosystem roles

Single-celled *Eucarya* are important members of Antarctic aquatic microbial communities. In many Antarctic systems, eucaryal algae are the main photosynthetic organisms and in others, only heterotrophic protists occupy the top trophic level. As eucaryal cells are generally large with characteristic morphologies, microscopic identifications have been used. However, microscopy is unable to classify smaller cells such as nanoflagellates with high resolution, although these may constitute a high proportion of algal biomass. For example, five morphotypes of *Chrysophyceae*, evident in Antarctic lakes were unidentifiable by light microscopy but were able to be classified using DGGE and DNA sequencing (Unrein *et al.*, 2005). Consistent with this, molecular studies specifically targeting eucaryal diversity (Unrein *et al.*, 2005; Mosier *et al.*, 2007; Bielewicz *et al.*, 2011) have identified a much higher level of diversity than previously suspected, and the studies have discovered lineages not previously known to be present such as silicoflagellates of the family *Dictyochophyceae* (Unrein *et al.*, 2005) and fungi (Mosier *et al.*, 2007; Bielewicz *et al.*, 2011).

Most *Eucarya* in Antarctic lakes are photosynthetic microalgae that are present in marine environments with a wide distribution including chlorophytes, haptophytes, cryptophytes and bacillariophytes. Molecular methods have afforded deeper insight into the phylogenetic diversity within these broader divisions and have revealed some patterns in their distribution. Using 18S rRNA gene amplification and DGGE, the same chrysophyte phylotypes were identified in lakes from the Antarctic Peninsula and King George

Island despite being 220 km apart (Unrein *et al.*, 2005) indicating these species may be well-adapted to Antarctica or highly dispersed. Similarly, an unknown stramenopile sequence was detected throughout the 18S rRNA clone libraries of Lake Bonney demonstrating a previously unrecognized taxon occupied the entire photic zone in the lake (Bielewicz *et al.*, 2011). In contrast, other groups showed distinct vertical and temporal distributions with cryptophytes dominating the surface, haptophytes the midwaters and chlorophytes the deeper layers during the summer while stramenopiles increased in the winter (Bielewicz *et al.*, 2011). Further studies are necessary to determine the basis for apparent specific adaptations of some species to particular lakes or lake strata, and for the cosmopolitan distribution of others. Here, molecular based research of the kind that has been applied to bacteria such as functional gene surveys will undoubtedly help answer these questions.

1.5 Limitations of taxonomic surveys

Inferring functional potential from taxonomic surveys can be problematic due to species or strain level differences in otherwise related bacteria. For example, the majority of the *Gammaproteobacteria* in hypersaline lakes were relatives of *Marinobacter* suggesting that this genus is particularly adapted to hypersaline systems (Bowman *et al.*, 2000b; Glatz *et al.*, 2006; Matsuzaki *et al.*, 2006; Mosier *et al.*, 2007). Nonetheless, *Marinobacter* species from different lakes appeared biochemically distinct as isolates from hypersaline lake Suribati-Ike were all able to respire dimethylsulphoxide (DMSO) but not nitrate (Matsuzaki *et al.*, 2006). In contrast, those from the west lobe of Lake Bonney were all able to respire nitrate (Ward and Priscu, 1997). Interestingly, in the east lobe of the same lake, nitrate respiration was inhibited although a near-identical *Marinobacter* phylotype was present; it was speculated that the inhibition may have been caused by an as yet unidentified chemical factor (Ward *et al.*, 2005; Glatz *et al.*, 2006).

This also applies to *Eucarya*, as the influence of flagellates on ecosystem function is not necessarily clear-cut as they can simultaneously inhabit several trophic levels. For instance, in Ace Lake the mixotrophic phytoflagellate *Pyramimonas gelidocola* derives a proportion of its carbon intake through bacterivory (Bell and Laybourn-Parry, 2003) but in the nearby Highway Lake, it uptakes dissolved organic carbon (Laybourn-Parry *et al.*, 2005). This again illustrates potential limitations for deriving ecosystem level functions from taxonomic studies alone, even with taxa that appear physiologically straightforward.

1.5.1 Functional gene studies of Antarctic lakes

Studies based on functional genes analyse the presence and diversity of gene(s) unique to a particular microbial process that serve as a marker for it. They are similar taxonomic studies but are interested in the diversity and distribution of a gene that represents metabolic potential.

1.5.2 Integrative studies to derive whole ecosystem function

1.6 ‘-omics’ approaches

Metagenomic studies have assessed both the taxonomic composition and genetic potential of lake communities, and in some cases have linked function to specific members of the community (López-Bueno *et al.*, 2009; Ng *et al.*, 2010; Lauro *et al.*, 2011; Yau *et al.*, 2011; Varin *et al.*, 2012). When coupled with functional “omic” techniques (to date metaproteomics has been applied, but not metatranscriptomics or stable isotope probing), information has also been gained about the genetic complement that has been expressed by the resident populations (Ng *et al.*, 2010; Lauro *et al.*, 2011). More detailed description of the advances of “-omics” approaches is detailed in the introduction of chapter 2.

Metagenomics has enabled unprecedented insight into viral diversity and ecology by permitting more precise classification, information on genetic content and discovery of novel species. Metagenomic analysis of the virome of the freshwater Lake Limnopolar, Livingston Island uncovered the greatest depth of viral diversity of any aquatic system to date (López-Bueno *et al.*, 2009). Representatives from 12 viral families were detected, but unlike the two previous viromes that had been published at that time using comparable techniques, ssDNA viruses and large dsDNA viruses that putatively infect *Eucarya* were the dominant viral types rather than bacteriophages. The ssDNA viruses were related to circoviruses, geminiviruses, nanoviruses and satellites; viruses previously only known to infect plants and animals indicating they are much more diverse than previously suspected and may constitute new viral families. Samples taken in summer showed a shift in the viral community composition towards phycodnaviruses similar to *Ostreococcus tauri* virus, OtV5. This shift potentially reflects an increase in the host algae that are stimulated to bloom by the increased light availability.

Understanding of Antarctic viral diversity and ecology is still in its early days as a complete viral survey is problematic due to the lack of a universal viral gene or even universal genetic material. Furthermore, the enormous depth of viral diversity remains largely unsampled so most viral sequences have no significant similarity to sequence data repositories (López-Bueno *et al.*, 2009). What is clear is that viruses perform a crucial role in shaping community structure, driving host evolution, contributing to the dissolved nutrient pool, and understanding them is essential to understanding ecosystem function (?).

1.7 Objectives

This study aimed to use a primarily metagenomic approach complemented microscopy, metagenomics and abiotic parameters to gain an integrative understanding of whole ecosystems. Using this methodology, not only can the taxonomic composition of the lakes be determined but also the functional potential of the microbial population and insight into the active members of the community. Ace Lake and Organic Lake were chosen as

the study sites because as meromictic systems, differences in the microbial population can be examined across vertical gradients, they are also largely isolated, as marine-derived systems, adaptation of marine microbiota to lake system can be examined, they are sites of moderate diversity and may be reservoirs of unknown taxa.

The objectives of the research were:

1. To develop complementary analyses to metagenomic analysis.
2. To determine the microbial and viral composition of the lake communities and their functional potential.
3. To identify the genes expressed by the community.
4. To reconstruct genomic information of dominant taxa of interest.
5. To integrate environmental and biological data to model the lake microbial interactions and geochemical processes.

Chapter 2

Development of quantitative epifluorescence microscopy and metaproteomic methods to complement metagenomic analysis of Ace Lake

Co-authorship statement

Sections from this chapter 2 have been published as:

Federico M. Lauro, Matthew Z. DeMaere, **Sheree Yau**, Mark V. Brown, Charmaine Ng, David Wilkins, Mark J. Raftery, John A.E. Gibson, Cynthia Andrews-Pfannkoch, Matthew Lewis, Jeffery M. Hoffman, Torsten Thomas, and Ricardo Cavicchioli. An integrative study of a meromictic lake ecosystem in Antarctica. *International Society of Microbial Ecology Journal* 5:879–895, 2011.

I performed the metaproteomic mass spectra analysis, epifluorescence imaging, microbial and viral counts and wrote the corresponding sections of the publication. Only these parts of the publication are included in the results and discussion of this chapter.

Analyses performed by others that support the work presented in this chapter are as follows: Research was designed by Federico Lauro, Mark Brown, Torsten Thomas, John Gibson and Ricardo Cavicchioli. Sample collection was performed by Federico Lauro, Mark Brown, Torsten Thomas, Jeffery Hoffman and Ricardo Cavicchioli. DNA extraction and clone library preparation of 2006 samples was performed by Cynthia Andrews-Pfannkoch and Jeffery Hoffman of the J.Craig Venter Institute (JCVI). DNA sequencing quality control was performed by Matthew Lewis of the JCVI. Metagenomic sequence filtering, mosaic assembly and annotation was performed by Matthew DeMaere. Protein extraction, one-dimensional sodium dodecyl sulphate-polyacrylamide gel electrophoresis

and liquid chromatography mass spectrometry performed by Charmaine Ng. Assistance in mass spectra analysis was provided by Mark Raftery.

2.1 Abstract

2.2 Introduction

Ace Lake is a meromictic saline lake in the Vestfold Hills, Antarctica. It is the best studied of all the meromictic lakes in the Vestfold Hills and possibly Antarctica. Extensive physical, chemical and biological data has been collected from Ace Lake in the last decades (?). However, the microbial community has largely been probed using culture-based or phenotypic means, or inferred from the lake's geochemistry. An analysis of the 16S small subunit ribosomal RNA (SSU) diversity was completed of the sediment showed microbial diversity was reduced (Bowman *et al.*, 2000b).

Ace Lake is a highly stratified lake system, 25 m deep at its deepest point. It is ice-covered for approximately 9 months of the year and thaws some summers. Water is marine-derived and a largely neutral water balance has ensured salinity is close to that of seawater. Although a lens of fresher water from melted surface ice can be generated in the summer months, this only mixes when the lake is ice-free and not to great depth (?). The lake is physically separated into an aerobic mixolimnion, a steep chemocline/oxycline at 12.7 m and an anoxic monimolimnion below. The monimolimnion is sulfidic and methanogenic; both compounds have presumably accumulated through activity from sulphate-reducing bacteria (SRB) and methanogenic archaea respectively.

As a physically and chemically well-characterised system of moderate diversity, Ace Lake was chosen as a model ecosystem to implement an integrated metagenomic and metaproteomic analysis. Combining both approaches would allow assessment of the metabolic potential of the system and identify the active members of the community and processes at time of sampling. In other words, discern who's there?, what could they be doing? and are they really doing it? Samples were obtained down the depth profile at 5, 11.5, 12.7, 18 and 23 m depths corresponding to each of the three zones. Sampling was conducted as part of the global ocean sampling (GOS) expedition (Rusch *et al.*, 2007) by using size fractionation of microbial biomass onto 3.0, 0.8 and 0.1 μm membrane filters.

Microbial and viral community composition was assessed from the metagenomic dataset. Significant differences were found in taxonomic composition of each size fraction within each sample depth and stratification of the microbial community between the three zones of Ace Lake. The mixolimnion community is similar to a marine surface water assemblage consisting of a high abundance of the SAR11 clade of *Alphaproteobacteria* related to "*Candidatus Pelagibacter ubique*" and green algae of the order *Mamiellales*. However, diversity is reduced by one order of magnitude reduced (Lauro *et al.*, 2011). Unlike Southern Ocean surface water, the mixolimnion is overrepresented in *Cyanobacteria* related to *Synechococcus* and *Actinobacteria*, which may represent taxa that mark the transition of a marine to lake community. A dense, near clonal population of green sulphur bacteria related to *Chlorobium* termed C-ace reside at the chemo/oxycline at 12.7 m (Ng *et al.*, 2010; Lauro *et al.*, 2011). Below, in the anoxic monimolimnion is a diverse, primarily heterotrophic community with abundant SRB and methanogenic

Archaea.

Preliminary work on the metaproteome down the vertical profile of Ace Lake has been performed using the National Center for Biotechnology Information (NCBI) non-redundant database (NR) database (?). However, there were few protein identifications were achieved using the unmatched genomic database (Ng, 2010). Identification rate reduced as diversity of the sample increased. A focused metaproteogeonomic analysis conducted on the dense green sulphur bacteria (GSB) layer using the matched metagenome as the database resulted in many more protein identifications compared to using the NR database (Ng, 2010). Assessment of the genetic complement of Ace Lake showed a concurrent stratification of the functional potential in each zone of Ace Lake (Lauro *et al.*, 2011). Significantly, GSB appeared to be crucial in the lake ecosystem as they had the greatest genetic potential for nitrogen and carbon fixation as well as sulphur cycling(Ng, 2010; Lauro *et al.*, 2011). Metaproteomic analysis was able to identify which proteins were actively expressed and thus the active pathways of the GSB metabolism which are so crucial to the function of the lake (Ng *et al.*, 2010).

This study aimed to expand on the metagenomic analysis of the water column of Ace Lake using complementary analyses. Metaproteomic analysis was used to identify expressed proteins of the 0.1 μm size fraction Ace Lake using a matched metagenomic databases for protein identification to infer which taxa and metabolic processes were active at time of sampling. To determine cellular and viral densities and validate the efficacy of size-fractionation with a modified epifluorescence microscopy procedure was developed and implemented.

2.3 Materials and methods

2.3.1 Ace Lake samples

Water samples were collected from Ace Lake ($68^{\circ}28'23.2''\text{S}$, $78^{\circ}11'20.8''\text{E}$), Vestfold Hills, Antarctica on 21 and 22 December 2006. A 2 m hole positioned above the deepest point (25 m depth) of the lake was drilled through the ice cover of Ace Lake to reach the lake surface. A volume of 1–10 L was collected by sequential size fractionation through a 20 μm pre-filter directly onto filters 3.0, 0.8 and 0.1 μm pore-sized, 293 mm polyethersulfone membrane filters (Rusch *et al.*, 2007), along the depth profile as described previously (Ng *et al.*, 2010). Samples were taken in the order, 23, 18, 14, 12.7, 5 and finally, 11.5 m.

After samples from each depth were collected, the sample racks were sequentially washed with 2×25 L 0.1 N NaOH, 2×25 L 0.053% NaOCl, and 2×25 L fresh water. The sample hose was flushed with water from each depth before being applied to the filters. A *Chlorobium* signature was identified at 5 m, but not immediately above the GSB layer at 11.5 m. As the next sample taken after sampling at 12.7 m was at 5 m, and then 11.5 m, despite all equipment being thoroughly washed with bleach, NaOH and water, the simplest explanation for the GSB signature at 5 m is carry-over from sampling of the dense biomass at 12.7 m.

A sonde probe (YSI model 6600, YSI Inc., Yellow Springs, OH, USA) was used to record depth, dissolved oxygen content, pH, salinity, temperature and turbidity throughout the water column of the lake. Total organic carbon was determined using a total organic carbon analyzer, TOC-5000A (Shimadzu, Kyoto, Japan) equipped with a ASI-5000A auto sampler (Shimadzu), and particulate organic carbon by standard protocols (<http://www.epa.gov/glnpo/lmmb/methods/about.html>) at the Centre for Water and Waste Technology, UNSW.

2.3.2 DNA sequencing and data cleanup

DNA extraction and Sanger sequencing was performed on 3730xl capillary sequencers (Applied Biosystems, Carlsbad, CA, USA) and pyrosequencing on GS20 FLX Titanium (Roche, Branford, CT, USA) at the J.Craig Venter Institute in Rockville, MD, USA (Rusch *et al.*, 2007). The scaffolds and annotations will be available via Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) and public sequence repositories such as the NCBI and the reads will be available via the NCBI Trace Archive.

Sanger reads were trimmed according to quality clear ranges. The quality of pyrosequencing reads was assessed as follows: a basic local alignment search tool (BLAST) nucleotide database was created from the Sanger reads of the 0.1 µm fraction of samples GS230, GS231 and GS232 (see Table of metagenomic data). After blasting the corresponding pyrosequencing reads against each database with a minimum bitscore of 80 and maximum e-value of 0.1, reads were binned according to length. The percentage of reads for each bin lacking a match to the Sanger read database was recorded. The percentage reads at least 25% repetitive after MDUST (?) analysis at default settings, and the percentage of reads containing N's, were assessed. In contrast to earlier pyrosequencers (?), no length-dependent bias in reads containing N's was observed. However, short reads had a disproportionately high number of repeats. Moreover, based on the proportion of reads with no match to the Sanger data set, both very short and very long reads had a disproportionately high number of errors; an observation that was previously reported (?).

On the basis of this analysis, a three step filtering process was applied to each sample. Reads were initially run through the Celera SFFToCA (?) pre-processor followed by LUCY (?) and finally, excluding the bottom 8% and top 3% of reads determined from the read length distribution. As the SFFToCA (v5.3) pre-processor removes all reads with a perfect prefix of any other read it overcomes the ‘perfect duplicates problem (?).

After this process, <5% of the reads belonged to clusters of duplicates with three or more reads, and clusters of orthologous groups (COG) of proteins classification of these reads showed an over-representation of category L (replication, recombination and repair) that includes mobile genetic elements, which are often duplicated, suggesting a potential biological significance for the duplicated reads. It is possible these residual duplications are a result of high gene copy number or localized fragility of DNA sequences that might be biasing the shear points.

2.3.3 Metagenomic DNA assembly and annotation

Mosaic assemblies were generated for each sample fraction using Celera whole genome shotgun (WGS) Assembler v5.3 (?). For each assembly, the runtime parameters used were as outlined for 454 sequencing data in the published standard operating procedure (http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=14SFF_SOP). As none of the samples can be considered clonal, these are regarded as stringent assemblies (Rusch *et al.*, 2007). Each 0.1 µm fraction assembly was a hybrid of Sanger and 454 read data, wherein the estimated genome size was manually set to minimize the number of unitigs from abundant organisms being falsely classified as degenerate (Rusch *et al.*, 2007). Annotation of each sample fraction assembly was carried out using an in-house pipeline, wherein the pipeline stages consisted of genomic feature detection and subsequent annotation. Detected features consisted of open-reading frames (ORFs), transfer RNA and ribosomal RNA (rRNA). Each detected ORF was further annotated by BLAST comparison against NR, Swissprot and Kyoto Encyclopedia of Genes and Genomes (KEGG)-peptide sequence databases and by biosequence analysis using profile hidden Markov models (HMMER) comparison against the Institute of Genomic Research curated protein database (TIGRFAM) (?), COG (??) and known marker genes (?). In all cases the cut-off e-value was a maximum of 1e-5.

2.3.4 Epifluorescence microscopy

Samples of unfiltered lake water and the flow-through from 3.0 and 0.8 µm filters from all depths were collected on November 2008 and fixed on site in formalin 1% (v/v). The samples were stored at -80°C for subsequent direct counts of cells and virus-like particles (VLPs). Enumeration was performed according to the method of Patel *et al.* (2007) with modifications. Lake water samples were filtered onto 0.01 µm pore-size polycarbonate filters (25 mm Poretics, GE Osmonics, Minnetonka, MN, USA). Filters were air dried, then placed with the back of the filter on top of a 30 ml aliquot of 0.1% (w/v) molten low-gelling-point agarose and allowed to dry at 30°C. Samples were stained by the addition of 1 ml working solution (1 in 400 dilution in 0.02 µm filtered sterile Milli-Q) of SYBR Gold (Molecular Probes, Eugene, OR, USA) to 25 ml of mounting medium (VECTASHIELD HardSet, Vector Laboratories, Burlingame, CA, USA). Stained samples were counted immediately, or stored at -20°C for up to a week before counting. Samples were visualised under wide-blue filter set (excitation 460–495 nm, emission 510–550 nm) with an epifluorescence microscope (Olympus BX61, Hamburg, Germany).

2.3.5 Metaproteomic analysis

Proteins were extracted from membrane filters from all 0.1 µm fractions from the six depths (5, 11.5, 12.7, 14, 18 and 23 m), and one dimensional-sodium dodecyl sulphate polyacrylamide gel electrophoresis (1D-SDS PAGE) and in gel trypsin digestion, liquid chromatography and mass spectrometry (MS), and two dimensional mass spec-

trometry (MS-MS) data analysis and validation of protein identifications performed as previously described (Ng *et al.*, 2010), with minor modifications. The spectra generated were searched against the protein sequence database corresponding to that depth constructed from the 0.1 μ m mosaic assemblies. The number of protein sequences in each database were as follows: 5 m, 138,208; 11.5 m, 133,948; 12.7 m, 27,142; 14 m, 62,436; 18 m, 71,512; and 23 m, 128,878. Scaffold (version Scaffold_2.05.01, Proteome Software Inc., Portland, OR, USA) was used to validate MS-MS-based peptide and protein identifications. Peptide and protein identifications were accepted if they could be established at >95% and 99% probability, respectively, as specified by the Peptide Prophet algorithm (?). Protein identifications required the identification of at least two peptides.

Proteins that contained similar peptides and could not be differentiated based on MS-MS analysis alone were grouped to satisfy the principles of parsimony and are referred to as a protein group. Spectral counting was used to semi-quantitatively estimate protein abundance. The total assigned spectra that matched to each identified protein were exported from SCAFFOLD 2.0. For similar proteins that have shared peptides (a protein ambiguity group), spectra were assigned to the protein with the most unique spectra. To normalize for variation in total spectra acquired between sample replicates, the number of spectra of each protein was multiplied by the average total spectra divided by the total spectra of the individual replicate. The spectral count of each protein was averaged across the replicates. As longer proteins are more likely to be detected, the average spectral counts were divided by the length of the protein. This value is equivalent to the normalized spectral abundance factor (??). In order to compare the relative abundance of proteins between depths, the normalized spectral abundance factor was divided by the average read depth of the contig (scaffold or degenerate) to which the protein mapped.

If >90% of a scaffolds length consisted of surrogate (highly degenerate unitig) sequence, the average read depth of the surrogate was used. For identified proteins that were part of a protein group the longest protein length and largest read depth value in the group was used. Pairwise comparisons of each zone were conducted on COG assigned proteins. The normalized spectral counts from each protein was aggregated based on their COG annotation. All proteins that were part of an ambiguity group were confirmed to share the same COG annotation to ensure counts were not biased because of the common spectra.

The summed spectral counts from 5 and 11.5 m (mixolimnion), and 14, 18 and 23 m (monimolimnion) were pooled. Statistical significance of differences between each zone was assessed using Fisher's exact test, with confidence intervals at 99% significance calculated by the NewcombeWilson method and Holm-Bonferroni correction (P-value cutoff of 1e-5) in Statistical Analysis of Metagenomic Profiles (STAMP) (?). All proteins identified, including their gene identifier, normalized spectral abundance, COG and KEGG Orthology identifiers, KEGG locus tag and matching COG or KEGG description are provided in Supplementary Table S1.

2.4 Results and discussion

2.4.1 Development of epifluorescence microscopy methodology for cell and VLPs enumeration

As part of the integrative study of Ace Lake, an epifluorescence microscopy methodology was developed. Motivation for visualising microbiota from size fractionated water samples was twofold. Firstly, visualising microbiota from water samples allows examination of cellular morphologies and more importantly, enumeration of cells and VLPs. Absolute cellular and VLP densities are not easily inferred from the metagenomic data and necessitates a complementary method of determination. Determining viral densities is extremely important as the first studies to do so found that viruses are the most numerous biological entities on the planet and likely play a large role in plankton mortality in the ocean (Bergh *et al.*, 1989; Proctor and Fuhrman, 1990). VLP counts from marine environments vary with depth and trophic status ranging from 10^6 to 10^8 VLP ml $^{-1}$ (?). Viral counts from various aquatic environments have differed from site to site indicating a variable role of virally mediated mortality.

Secondly, size fractionation of suspended microbial biomass from aquatic environments has been utilised as part of the landmark Sargasso Sea metagenomic study (Venter *et al.*, 2004) and subsequent GOS expedition (Rusch *et al.*, 2007). The Ace Lake samples were collected using the same sampling strategy as the GOS dataset but has sequence information from all three filter sizes. Thus, the visualisation of microbial morphologies from each stage of the filtration process would allow validation of the filtration process.

Developing a revised method for the simultaneous counts of cells and VLPs was necessary due to inability to source 25 mm diameter 0.02 μm pore-size Anodisc filters (Whatman) that have been long used with fluorescent nucleic acid dyes for this purpose (?Noble and Fuhrman, 1998). They were marked for discontinuation in December 2008 after the take over of Whatman by GE Healthcare and was the cause of a global shortage that was strongly opposed by the viral ecology community (Torrice, 2009). Since conducting this research, production of 25 mm Anodisc filters has resumed after the petitioning of GE Healthcare by the environmental virology community. Cost of Anodiscs filters are relatively higher than they were previously. Nonetheless, the need for alternative methodologies during the shortage was so great that alternative protocols were developed independently in other research groups (Budinoff *et al.*, 2011; Diemer *et al.*, 2012) stressing the importance of an affordable method of direct viral counts and the utility of alternative methodologies.

Clear polycarbonate Track EtchTM (PCTE) filters were selected for use as a viable alternative product for the following reasons.

1. They have defined pore-size.
2. They are available in 0.015 μm pore-size allowing capture of VLPs.
3. They have previously been used for VLPs enumeration (??).

4. They have a long history of use with cellular enumeration (?) and can therefore be easily adopted for widespread use.

There are several disadvantages to using PCTE membranes over Anodiscs which are as follows.

1. Polycarbonate is not as robust as 25 mm diameter Anodisc filters that have a built in plastic support ring around the edge.
2. The polycarbonate of PCTE filters cannot be blotted dry as the Al_2O_e of Anodisc filters.
3. PCTE filters appear to have high background fluorescence.
4. PCTE filters have a much slower flow rate than Anodisc filters.

The protocol used is detailed in the materials and methods 2.3. The main disadvantages of using PCTE were overcome for the purposes required for this study. To counteract the fragility of the PCTE filters, vacuum filtration was not performed above X pressure limit and filters. No cracks or tears were observed during visualisation. PCTE filters also have a tendency to crinkle when mounted on the glass slide making visualisation of cells and VLPs on a single focal plane difficult. Agarose was used to embed the filters to help flatten the membrane and aid in mounting. However, this was not strictly necessary if filters were dried well and the membrane mounted carefully so that it was pressed flat against the glass slide. The background fluorescence of the clear filters was low when stain is only incorporated into the mountant after filtration rather than staining in the column before filtration.

Filtration onto very small pore-size also necessitated a very strong seal of the filter column against the glass, but the slow flow-rate was not a property of the PCTE filters that could be overcome. As the filtration and visualisation was performed on fixed samples in the laboratory rather than in the field, the time taken for filtration of 2–3 hours for each sample, this was not deemed problematic for this study.

Overall, a viable alternative methodology was successfully developed for visualisation and enumeration of cells and VLPs using PCTE membrane filters was successful and could be used as an alternative to Anodisc-filters (Figure 2.1). To be competitive alternative to Anodisc filters in terms of accuracy of counts, a comparison of this methodology and Anodisc-based protocols using viral samples of known densities needs to be performed.

No Anodisc filters could be obtained for such a comparative analysis to validate its use in this study. However, for the purposes of this study, which is to show the relative differences in morphotypes between sample depths and size fractions, this method was more than suitable. Other groups have since shown it to be comparable.

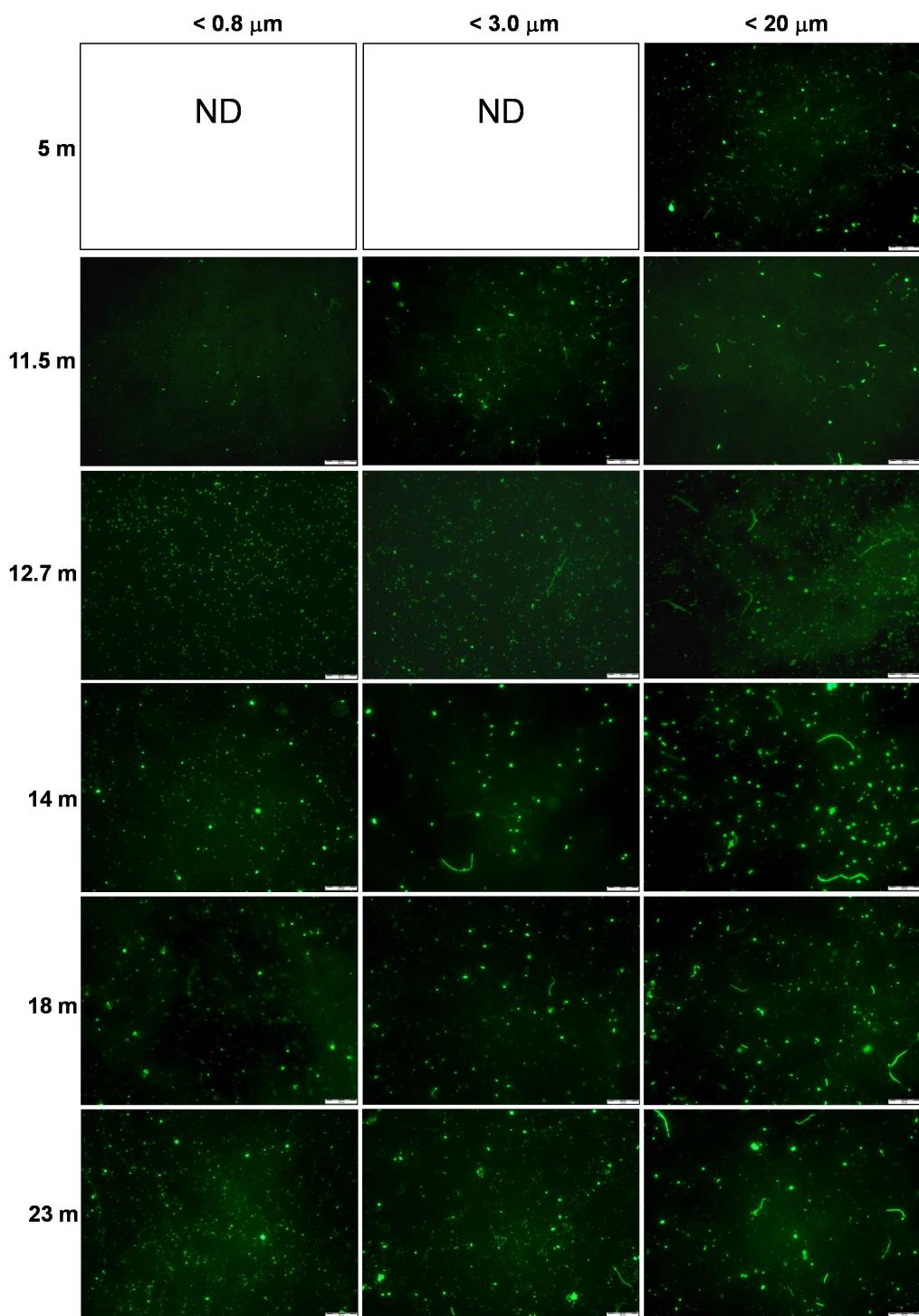


Figure 2.1: Epifluorescence microscopy images of Ace Lake microbiota. Scale bar = 20 μm . ND, not determined.

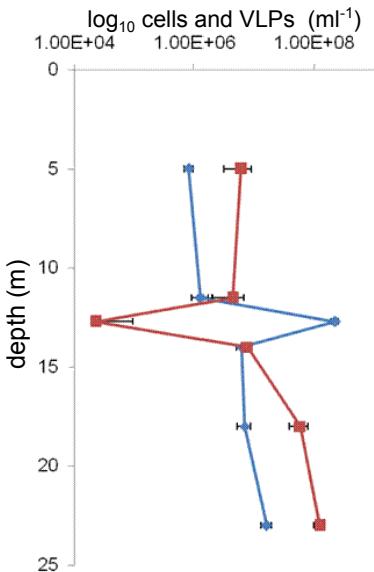


Figure 2.2: Counts of microbial cells and VLPs by epifluorescence microscopy down the water column of Ace Lake. Error bars represent one standard deviation. No VLPs were detected at 12.7 m depth and the value reported represent the detection limit of the counting procedure (*i.e.* one VLP detected in one field of view); the true number is likely to be lower than this.

2.4.2 Size and depth stratification of the community supported by cell and VLP densities

Development of fluorescence microscopy methodology using 0.01 μm pore-size polycarbonate filters for simultaneous cellular and viral counts shows:

1. Size fractionation procedure appeared effective.
2. Morphological differences supports stratification of the community.
3. Visualisation of the morphology supported the metagenomic data that saw size fractionated and taxonomically stratified community.
4. Virus to bacteria ratios can give important information about the community.

At 12.7m depth, the light levels, and the sharp transition in oxygen content and salinity (Fig. S2) favour the dominance of a very high-density ($2.2 \times 10^8 \text{ cells ml}^{-1}$) of a single type of GSB of the genus *Chlorobia*, referred to as C-Ace (Ng *et al.*, 2010). Viral signatures were essentially devoid in this zone. The ratio of bacteriophage to total viral population increased proportionally in the larger size fractions consistent with trophic analyses that indicate that the larger size fractions are mostly copiotrophic (Fig. S8) particle attached bacteria and therefore likely to be sensitive to lysogenic phage infection (Lauro *et al.*, 2009). The 23 m unfiltered lake water contained very high levels ($1.3 \times 10^8 \text{ VLPs ml}^{-1}$) of VLPs. The high diversity of bacteria and archaea in all size fractions of the monimolimnion (Fig. 2) is consistent with the presence of a high viral population (Rodriguez-Valera *et al.*, NRM, 2009).

2.4.3 Development of metaproteomic methodology

1. Using a matched metagenome instead of NR for protein identification greatly increased the number of identifications. 1.1 Except at the bottom zone, likely because the community is too diverse so greater coverage of the metagenome is required. In parallel with taxonomic diversity increasing with depth (with the exception of the GSB layer), the rate of metaproteomic identification of proteins decreased with depth (Table S2). The majority of the proteins that were detected (e.g. 67% at 23 m) were for hypothetical proteins that tended to lack orthologues in well-characterized organisms, highlighting both the functional importance and novelty of this anaerobic zone of the lake.

2. More specific information could be assigned to the taxonomic groups such as. 2.1. The Actinobacteria sequences in the mixolimnion were associated with a diverse phylogenetic cluster (Luna cluster) mainly contributed by freshwater ultramicrobacteria (?). Several Luna cluster isolates contain rhodopsin genes (Sharma *et al.*, 2009) and similar gene sequences were present in the Ace Lake oxic zone data and found to be expressed (167820670 and 163154474; Table S2). 2.2. This is consistent with the identification of clustered regularly interspaced short palindromic repeat (CRISPR) associated CRISPR-associated proteins (CAS) proteins Cse2, Cse3 and Cse4 (165526330, 165526332 and 165526334, respectively) in the 12.7 m metaproteome (Table S2). The CAS gene locus (cas3, cse1, cse2, cse3, cse4, cas5, cas1b), to which the proteins map, shares its organisation with CAS loci of sequenced GSB, and groups with the *E. coli* subtype/variant 2. The CRISPR/CAS system is likely to confer phage resistance to C-Ace, akin to the role in other organisms (Karginov and Hannon 2010; Horvath and Barrangou 2010).

3. Using SCAFFOLD to validate protein identification and perform spectral counts was helpful. 3.1 Same protein identifications as Charmaine except one or two. 3.2 Able to quantify differences between mixolimnion and monimolimnion. The diversity and abundance of ATP-binding cassette (ABC) transporters was lowest in the 0.1 µm fractions at 23 m (Fig. 3), and a correspondingly low number were detected in the metaproteome (Table S2). In contrast, numerous transporters, predominately ABC type, were identified in the metaproteome of the mixolimnion samples, with a high COG representation of transporters for carbohydrates (~34% of normalized spectra), amino acids (~32%) and inorganic ions (~9%) (Table S2 and Fig. S11). All transporters in the metaproteome were of bacterial origin and conservative phylum level assignments of the normalised spectra showed the majority to originate from *Proteobacteria* (69%), of which SAR11 comprised 46% and *Actinobacteria* 19% (Table S2). A high proportion of expressed genes with transport functions have also been reported for SAR11 from coastal (Poretsky *et al.* 2010) and open ocean waters (Sowell *et al.* 2009) (Morris *et al.* 2010?). Oligotrophs, such as SAR11 not only posses a low-diversity of high-affinity transporters (Lauro *et al.*, 2009), but regulate the relative abundance of transporters expressed in response to dissolved organic carbon (DOC) availability (Poretsky *et al.* 2010). The prevalence of amino acid and simple sugar transporters (Table S2), and the low DOC concentration in the Ace Lake mixolimnion (Fig. 1) is likely to reflect efficient

utilisation of these substrates from the DOC pool. Two SAR11 transport proteins that were detected in Ace Lake (Table S2) were not detected from the Sargasso Sea (Sowell,et al. 2009): an ectoine/hydroxyectoine (167807477 and 167892279) and a zinc ABC transporter (167933120). The zinc ABC transporter is likely to support zinc efflux in response to zinc concentrations which are ~70-fold higher in the mixolimnion of Ace Lake compared to seawater (Rankin 1999). Conversely, phosphate transporters were a major class detected from the Sargasso Sea (Sowell,et al. 2009) but were absent from the Ace Lake metaproteome; consistent with lower phosphate levels in the Sargasso Sea (<5 nM) compared to Ace Lake (1–12 µM). The differences in transporter expression between Ace Lake and oceanic SAR11 are likely to signify adaptive growth strategies that have evolved in the Ace Lake SAR11 community. The high numbers of bacteriophages in the monimolimnion (detected by microscopy, Fig. S5 and S6; metaproteomics, Table S2; metagenomics, Fig. 2), and increase in DOC observed at depth (Fig. 1), also indicates that carbon turnover in the monimolimnion is likely to be tightly coupled to the carbon flux going through a viral shunt, as proposed for open ocean systems (Suttle, C. A. Viruses in the sea. *Nature* 437, 356361 (2005)). The bacteriophages are also likely vehicles for mediating gene exchange. Most of the genetic potential to cycle the nitrogen pool appears to be limited to nitrogen assimilation throughout the lake and remineralisation in the monimolimnion (Fig. S14). The detection of glutamine synthetase (GlnA) and glutamate synthases (GltBS) in the metaproteome (Table S2) are supportive of active nitrogen assimilation. In the mixolimnion, GlnA was linked to SAR11 and *Actinobacteria*, and they are likely to be responsible for nitrogen absorption in the oxic zone. At the oxycline, GlnA and GltB from GSB were abundant (Table S2), indicating an important role for nitrogen assimilation at this zone in the lake. Genes for **ASR!** (**ASR!**) were present in metagenome data of all three fractions at all depths, although they were lowest at the oxycline. However, there was no evidence for expression of the genes as ASR proteins were not detected by metaproteomics. In contrast, multiple subunits of the GSB dissimilatory sulfide reductase complex were identified (Ng et al. 2010 and Table S2) indicating functionality of this pathway at the oxycline. GSB likely utilise the dissimilatory sulfide reductase system to convert sulphur to sulfite and the polysulfide-reductase-like complex 3 to oxidize sulfite to sulphate. SRB may then reform sulphide completing the sulphur cycle between the GSB and SRB (Ng et al. 2010). While SRB were detected at the three depths of the monimolimnion, sulphate is depleted in the water column and sediment at the bottom of the lake limiting their dissimilatory capacity (?). Finally, sulphate in the mixolimnion can be linked to sulphur-oxidation by SAR11 (Meyer and Kuever, *Microbiology* 153:3478–3498) and a concomitant lack of capacity to perform sulphur reduction.

2.5 Conclusions

Using complementary approaches helps to validate the research methodology and metagenomic inferences about the whole community. Specifically, differences in size and depth was shown by both microscopy and metagenomics to be apparent. This both validates

the method of size fractionation as a viable approach to broad separation of the community, as well as supports the assertion that there was a large difference in community at different depths. Using a matched metaproteomic database showed a huge increase in the number of protein identifications. This was provided that metagenomic coverage was good. Using a metaproteomics, genes identified as potentially relevant in the metagenome were found to be expressed, supporting their importance. For example, it showed the CRISPR genes were active and may be a defence against phage. It also showed Actinorhodopsins were expressed. It showed that abundant genes were normally abundant in the metaproteome, such as transport proteins that give insight into what substrates are important components of the DOC pool. New inferences could be drawn from the metaproteome, such as the preference for labile substrates such as active sulfate reduction which is not apparent from the sulphate concentration.

Chapter 3

Virophage control of Antarctic algal host–virus dynamics

Co-authorship statement

A version of this chapter has been published as:

Sheree Yau, Federico M. Lauro, Matthew Z. DeMaere, Mark V. Brown, Torsten Thomas, Mark J. Raftery, Cynthia Andrews-Pfannkoch, Matthew Lewis, Jeffery M. Hoffman, John A. Gibson, and Ricardo Cavicchioli. Virophage control of antarctic algal host–virus dynamics.

Proceedings of the National Academy of Sciences USA 108:6163–6168, 2011.

Contributions to this publication by other researchers is as follows. Research was designed by Federico Lauro, Mark Brown, Torsten Thomas, John Gibson and Ricardo Cavicchioli. Sample collection was performed by Federico Lauro, Mark Brown, Torsten Thomas, Jeffery Hoffman and Ricardo Cavicchioli. DNA extraction and clone library preparation of 2006 samples was performed by Cynthia Andrews-Pfannkoch and Jeffery Hoffman of the J. Craig Venter Institute. DNA sequencing quality control was performed by Matthew Lewis of the J. Craig Venter Institute. Metagenomic sequence filtering, global assembly and annotation was performed by Matthew DeMaere. Assistance in mass spectrometry and mass spectra analysis was provided by Mark Raftery. Assistance in analysis of Eucarya taxonomy was provided by Mark Brown. Analysis of virophage abundance over time was performed by Federico Lauro.

Apart from these contributions, I performed all other data analyses and interpretations.

3.1 Abstract

Viruses are abundant ubiquitous members of microbial communities, and in the marine environment affect population structure and nutrient cycling by infecting and lysing primary producers. Antarctic lakes are microbially dominated ecosystems supporting truncated food webs where viruses exert a major influence on the microbial loop. Here we report the discovery of a new virophage (relative of the recently described Sputnik virophage) that preys on phycodnaviruses that infect prasinophytes (phototrophic algae). By performing metaproteogenomic analysis on samples from Organic Lake, a hypersaline meromictic lake in Antarctica, complete virophage and near-complete phycodnavirus genomes were obtained. By introducing the virophage as an additional predator of a predator-prey dynamic model we determine that the virophage stimulates secondary production through the microbial loop by reducing overall mortality of the host and increasing the frequency of blooms during polar summer light periods. Virophages remained abundant in the lake two years later, and were represented by populations with a high level of major capsid protein sequence variation (25–100% identity). Virophage signatures were also found in neighbouring Ace Lake (in abundance), and in two tropical lakes (hypersaline and fresh), an estuary, and an ocean upwelling site. These findings indicate that virophages regulate host–virus interactions and influence overall carbon flux in Organic Lake, and play previously unrecognised roles in diverse aquatic ecosystems.

3.2 Introduction

It has been known for at least 20 years that viruses frequently infect and lyse marine primary producers causing up to 70% of cyanobacterial mortality (Proctor and Fuhrman, 1990; ?). Eucaryotic phytoplankton are preyed upon by large dsDNA phycodnaviruses (PVs) causing bloom termination in globally distributed species (3,6). Elevated levels of dissolved organic carbon (DOC) (7) and numbers of heterotrophic bacteria (8-10) occur during algal blooms indicating that viral lysis of eucaryotic algae stimulates secondary production. Viruses also suppress host populations at concentrations below bloom-forming levels, with abundance being controlled by the efficiency and production rates of the infecting viruses (11, 12). Antarctic lakes are microbially dominated ecosystems supporting few, if any metazoans in the water column (13, 14). In these truncated food webs, viruses are expected to play an increased role in the microbial loop (15). Low complexity Antarctic lake systems are amenable to whole community based molecular analyses where the role that viruses play in microbial dynamics can be unravelled (14). Attesting to this, a metagenomic study of Lake Limnopolar, West Antarctica uncovered a dominance of eucaryotic viruses and ssDNA viruses previously unknown in aquatic systems (16).

We established a metaproteogenomic program for Organic Lake, which is located in the Vestfold Hills, East Antarctica, in order to functionally characterize its microbial community. Organic Lake is a shallow (~ 7 m) hypersaline (~ 230 g L $^{-1}$ maximum salinity) meromictic lake with a high concentration of dimethylsulphide (DMS) (~ 120 μg^{-1}) in its monimolimnion (17, 18). Water temperature at the surface of the lake can vary from -14 to $+15^\circ\text{C}$ while remaining sub-zero at depth (19, 20). The lake is eutrophic, with organic material sourced both from autochthonous production and input from penguins and terrestrial algae. The high concentrations of organic material reflect slow breakdown in the highly saline lake water. The salt in the lake was trapped along with the marine biota when the lake was formed due to falling sea level $\sim 3,000$ BP (21, 22). The lake sediment has both low species diversity (Shannon-Weaver diversity: 1.01) and richness (Chao non-parametric index: 32 ± 12) (23). Unlike high latitude lakes, viral abundance has been reported to increase with trophic status (15) and with salinity in Antarctic lakes (24).

Here we report the analysis of the surface water of Organic Lake, highlighting the presence of a relative of the recently described Sputnik virophage, a small eucaryotic virus that requires a helper *Acanthamoeba polyphaga* mimivirus (ApMV) to replicate (25). From metagenomic DNA, a complete Organic Lake virophage (OLV) genome was constructed (the second virophage genome to be described), and near-complete genomes of its probable helper Organic Lake phycodnavirus (OLPV).

Table 3.1: Summary of metagenomic data for Organic Lake 0.1 μm fraction samples used in this study. SCF, scaffolds.

ID	Date	Trimmed reads	SCFs >10 kbp (reads in SCFs)	Annotated ORFs in SCFs (total ORFs)
GS233	December 2006	418,265 (Sanger: 28,481) (454: 389,784)	45 (221,573)	7,318 (21,961)
GS374	November 2008	494,573	5 (771)	33,262 (83,684)
GS379	December 2008	446,200	2 (40,314)	23,012 (64,779)

3.3 Materials and methods

3.3.1 Samples and DNA sequencing

Water samples collected from Organic Lake were: 1) Surface water from the eastern side of the ice-free lake ($68^{\circ}27'25.48''\text{S}$, $78^{\circ}11'28.06''\text{E}$) December 24, 2006. 2) A depth profile collected through a 30 cm hole drilled through the surface ice above the deepest point in the lake ($68^{\circ}27'22.15''\text{S}$, $78^{\circ}11'23.95''\text{E}$), November 10, 2008. 3) Surface water from the north-east side of the partially ice-covered lake ($68^{\circ}27'21.02''\text{S}$, $78^{\circ}11'42.42''\text{E}$), December 12, 2008. Samples were sequentially filtered through a 20 μm pre-filter and biomass captured onto 3.0, 0.8 and 0.1 μm membrane filters as described previously (1, 2). The samples from 2008 also included 50% (v/v) RNAlater. DNA extraction, sequencing and quality validation was performed as previously described (1, 2). DNA sequencing was performed at the J. Craig Venter Institute in Rockville, MD, USA.

3.3.2 Transmission electron microscopy

Unfiltered Organic Lake surface water from December 24, 2006 (fixed on-site in 1% (v/v) formalin) was concentrated and a solvent exchange performed with sterile filtered ammonium acetate solution 1% (w/v) using a 50 kDa cut-off Microcon centrifugal filter device (Millipore) according to the manufacturers instructions. Formvar coated 200 mesh copper grids were floated on a droplet of sample for 30 min, excess liquid wicked off and the grid negatively stained for 30 s with uranyl acetate 2% (w/v). The sample was visualised using a JEOL1400 transmission electron microscope at 100 kV at 150,000 to 250,000 \times magnification.

3.3.3 Metagenomic assembly and annotation

Mosaic metagenomic assemblies were generated as previously described (1, 2). For the 0.1 μm Organic Lake 2006 sample, assembly was a hybrid of Sanger and 454 read data (Table 3.1). For all other sample size fractions, runtime parameters used were standard for 454 sequencing data. Low GC ($\geq 51\%$) scaffolds >10 kbp from the 0.1 μm 2006 assembly had high coverage ($>45\times$) indicating these were from the dominant taxa. One of these scaffolds was binned as virophage and the rest as PV.

To further separate the OLPV types and assess the completeness of their genomic

Table 3.2: List of primers used to close the OLV genome. Dir, direction.

Primer function	ID	Dir.	Sequence (5'-3')	Length (bp)
Outer gap spanning	SY11	forward	TTG TCT TAT GTA TTA CAA ATC ATT GAA	3,843
Outer gap spanning	SY12	reverse	CGA CAT TAA TCG GTT GTT TT	
Nested gap spanning	SY13	forward	GCA TTA CGA ATG TGT TCC AG	3,403
Nested gap spanning	SY14	reverse	TTC TCC GTG ATT GAT ATC GT	
Sequencing	SY23	forward	TCC CTA TTG ATG TCA AAA CC	-
Sequencing	SY24	forward	GAT TCT GGT TGG AGC ATA TAT TT	

content, highly conserved single copy PV orthologues were identified as follows. An all against all basic local alignment search tool (BLAST)p search was conducted with protein sequences from the ten available PV genomes (*Acanthocystis turfacea Chlorella* virus 1, PbCV-1, PbCV AR158, PbCV FR483, PbCV NY2A, *Emiliania huxleyi* virus 86, *Ectocarpus siliculosus* virus 1, *Feldmannia* sp. virus, *Ostreococcus* virus 5, *Ostreococcus tauri* virus 1) and ApMV (which was included as a close PV relative). BLASTp results were parsed and clustered using ORTHOMCL V1.4 (3, 4).

Pairs of each orthologue were located on eight of the PV scaffolds. The location of each orthologue pair had a complementary distribution so the eight scaffolds were able to be sorted unambiguously into two strains (OLPV-1 and OLPV-2). OLPV-1 ribonuclease reductase α -subunit appeared as duplicated on different scaffold ends, likely as an artefact of its proximity to an assembly break point. The remaining high coverage scaffolds were searched for predicted proteins present in one OLPV strain but not in the other and assigned to the strain in which it was absent. Comparison of OLPV-1 and OLPV-2 scaffolds was performed using TBLASTN of concatenated scaffolds from each strain and visualised using the Artemis comparison tool (ACT) (5). DNA sequence data is available in Genbank and Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) (<http://web.camera.calit2.net>).

3.3.4 Organic Lake virophage genome completion and annotation

The high coverage (77 \times), large number of Sputnik homologues that encode essential functions and length of the putative OLV scaffold from the 0.1 μ m 2006 hybrid assembly indicated it was a near-complete genome. Reads from this scaffold were reassembled at high stringency and visualised using PHRED/PHRAP/CONSED (6) to complete the sequence. Mate-pair data indicated a circular molecule and primers were designed to span the ends of the scaffold and sequence across the gap (Table 3.2).

Touch-down polymerase chain reaction (PCR) was performed with DNA from 0.1 μ m 2006 sample, the product used for nested PCR and the final product was cloned and sequenced. The complete genome was manually annotated and visualised using ARTEMIS (7). Translated ORFs (minimal size 120 amino acids) were compared (BLASTp) to GenBank, to the all metagenomic ORF peptide database on CAMERA (<http://web.camera.calit2.net>) and to predicted proteins from OLPV-1 and OLPV-2 scaffolds. Comparisons between the OLV genome and OLPV-1/OLPV-2 were performed with

TBLASTN and visualised using ACT (5).

3.3.5 Phylogenetic analysis

Translated amino acid sequences from viral marker genes of interest were retrieved from the 0.1 μm 2006 metagenomic assemblies from this study, GenBank and CAMERA all metagenomic reads ORF peptide database. Homologous sequences were aligned using MUSCLE v3.6 (8). Neighbour-joining analysis, test for clade support (bootstrap analysis 2000 replicates) and tree drawing was performed with Molecular Evolutionary Genetic Analysis (MEGA) software v4 (9). Maximum likelihood analysis (JTT substitution model) and test for clade support (aLRT analysis) was performed with PHYLML (10) and the tree visualised using MEGA. 18S ribosomal RNA (rRNA) gene sequences were retrieved from reads of all filter sizes, compared (BLASTN, e-value <1.0e -5) to the SILVA100 SSURef database, aligned and phylogeny performed using ARB as previously described (1, 2). The abundance and similarity of virophages in all lake samples and filter sizes was estimated using BLASTP (evalue <1.0e -5) to search using the OLV major capsid protein (MCP) sequence against a database of proteins predicted from sequencing reads. The database was generated as previously described (1) and the percent identity of the BLAST hit was used as a proxy for species similarity.

3.3.6 Metaproteomic analysis

Metaproteomics of proteins from the 0.1 μm filter from 2006 was performed as previously described (1, 2), with minor modifications. The protein sequence database was generated by combining ORFs from the 3.0, 0.8 and 0.1 μm mosaic assemblies with 130,581 sequences in the database. SCAFFOLD 3.0 (Proteome Software Inc.) was used to validate MS/MS based peptide and protein identifications. The peptide sequences by which the proteins were identified shown in Appendix Table 6.1.

3.3.7 Model of algal host–virus and virophage dynamics

To model the effect a virophage would have on algal *Pyramimonas* algal host populations in Organic Lake, modified Lotka-Volterra equations were used describing the OLV as a predator of predator OLPV. The original equations are given by:

$$\frac{dA}{dt} = \alpha A - \varepsilon PA \quad (3.1)$$

$$\frac{dP}{dt} = \theta PA - \mu P \quad (3.2)$$

Where:

A is the number of *Pyramimonas* (prey).

P is the number of OLPV (predator).

α is the specific growth rate of the prey.

θ is the specific production rate of the predator.

ε is the rate of predator mediated death of prey.

μ is the specific decay rate of the predator.

Equation 3.1 describes the change in *Pyramimonas* abundance and equation 3.2 the change in OLPV abundance in the absence of OLV. In the presence of OLV, *Pyramimonas*, OLPV and OLV dynamics are described by the following equations:

$$\frac{dP}{dt} = \theta PA - \mu P - \omega PV \quad (3.3)$$

$$\frac{dV}{dt} = \beta PV - \gamma V \quad (3.4)$$

Where:

V is the number of the OLV (predator of predator).

ω is the rate of OLV mediated reduction in OLPV infective particles.

β is the production rate of OLV.

γ is the decay rate of OLV.

Equation 3.3 is a modified version of equation 3.2 which includes the effect of OLV on the change in abundance of OLPV. Equation 3.4 describes the growth properties of OLV as a predator of OLPV. Values for the variables for the solution shown (Figure 3.9) were as follows: initial prey (10), predator (1) and predator of predator (10) numbers, $\alpha = 0.1$, $\theta = 0.0015$, $\varepsilon = 0.01$, $\mu = 0.01$, $\omega = 0.01$, $\beta = 0.015$ and $\gamma = 0.15$. COmplex PAthway Simulator (COPASI) software (11) was used to simulate prey, predator and predator of predator dynamics using the deterministic (LSODA) method.

3.4 Results and discussion

3.4.1 Dominance of phycodnaviruses in Organic Lake

Water samples from Organic Lake were collected December 2006 and November and December 2008, and microbial biomass collected onto 3.0, 0.8 and 0.1 μm membrane filters as described previously (14). A large proportion of shotgun sequencing reads (96.2%) from the 0.1 μm size fraction of the 2006 Organic Lake metagenome (Table 3.1) had no significant hits to sequences in the RefSeq database (TBLASTx with e-value <1.0e-3, minimum alignment length: 60 bp, minimum identity: 60%). The degree of assembly was high, with 77% of reads forming part of a scaffold, indicating the sample contained a few abundant taxa of minimal diversity. Forty-five scaffolds were longer than 10 kbp; the five longest ranged from 70 to 171 kbp. GC content and coverage were used to separate scaffolds into taxonomic groups (Figure 3.1). A broad division was evident between low ($\leq 41\%$) and high ($\geq 51\%$) GC scaffolds suggesting they constituted two taxonomic groups. All scaffolds in the high GC group that could be assigned contained

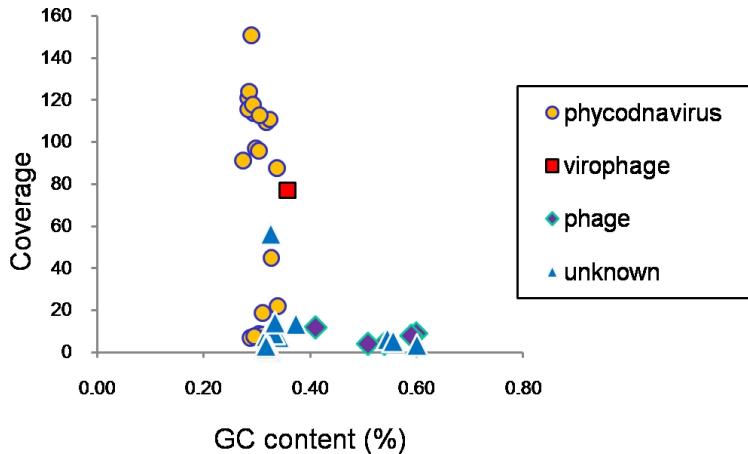


Figure 3.1: Plot of percent GC content vs coverage for the 2006 Organic Lake 0.1 μm hybrid assembly scaffolds >10 kb.

phage homologues, as did the one exceptional low GC scaffold. The low coverage in the high GC group showed bacteriophages were not abundant in the 0.1 μm fraction. These scaffolds were not analyzed further. The low GC scaffolds with confident assignments contained sequences matching conserved phycodnavirus (PV) or *Acanthamoeba polyphaga* mimivirus (ApMV) proteins. These PV-related scaffolds comprised 60% of assembled reads demonstrating that Organic Lake phycodnaviruses (OLPVs) were numerically dominant in the 0.1 μm fraction. transmission electron microscopy (TEM) revealed the presence of virus-like particles with the dimensions and structure typical of PVs (Figure 3.2A).

Within the low GC group, scaffolds separated into a high coverage ($>45\times$) group, including the five longest scaffolds, and a low coverage ($<22\times$) group. Two of the scaffolds in the high coverage group and one in the low coverage group contained the PV marker DNA polymerase B (DPOB). The two high coverage DPOB share 76% amino acid identity and both share $\sim 57\%$ identity to the low coverage DPOB. DPOB is single-copy throughout the nucleo-cytoplasmic large DNA virus (NCLDV) family to which PVs belong (26,27), demonstrating that the Organic Lake surface waters contained two closely related abundant PV types (DPOB1) and (DPOB2), and a more distantly related lower abundance type (DPOB3).

Phylogenetic analysis clustered Organic Lake DPOB with unclassified lytic marine PV isolates that infect the prymnesiophytes *Chrysochromulina ericina* (CeV1) and *Phaeocystis pouchetii* (PpV), the prasinophyte *Pyramimonas orientalis* (PoV) (4,28), and uncultured marine PVs related to ApMV (29, 30) (Figure 3.3). As the host range of PVs broadly correlates with DPOB phylogeny (31, 32), OLPVs would infect prasinophytes or prymnesiophytes. The most probable host is the prasinophyte, *Pyramimonas* as no prymnesiophyte 18S ribosomal RNA (rRNA) gene sequences were present in any size fraction of the Organic Lake metagenome (Figure 3.4).

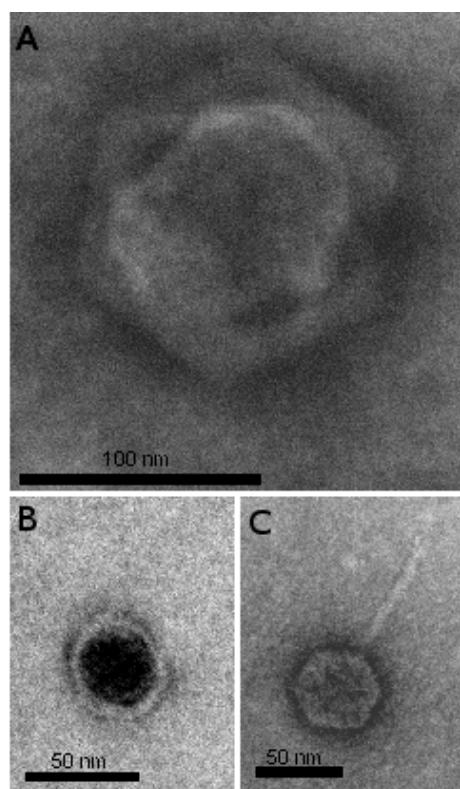


Figure 3.2: Transmission electron micrographs of negatively stained virus-like particles from Organic Lake. **(A)** VLP resembling the size and morphology of PVs, **(B)** Sputnik virophage and **(C)** bacteriophages.

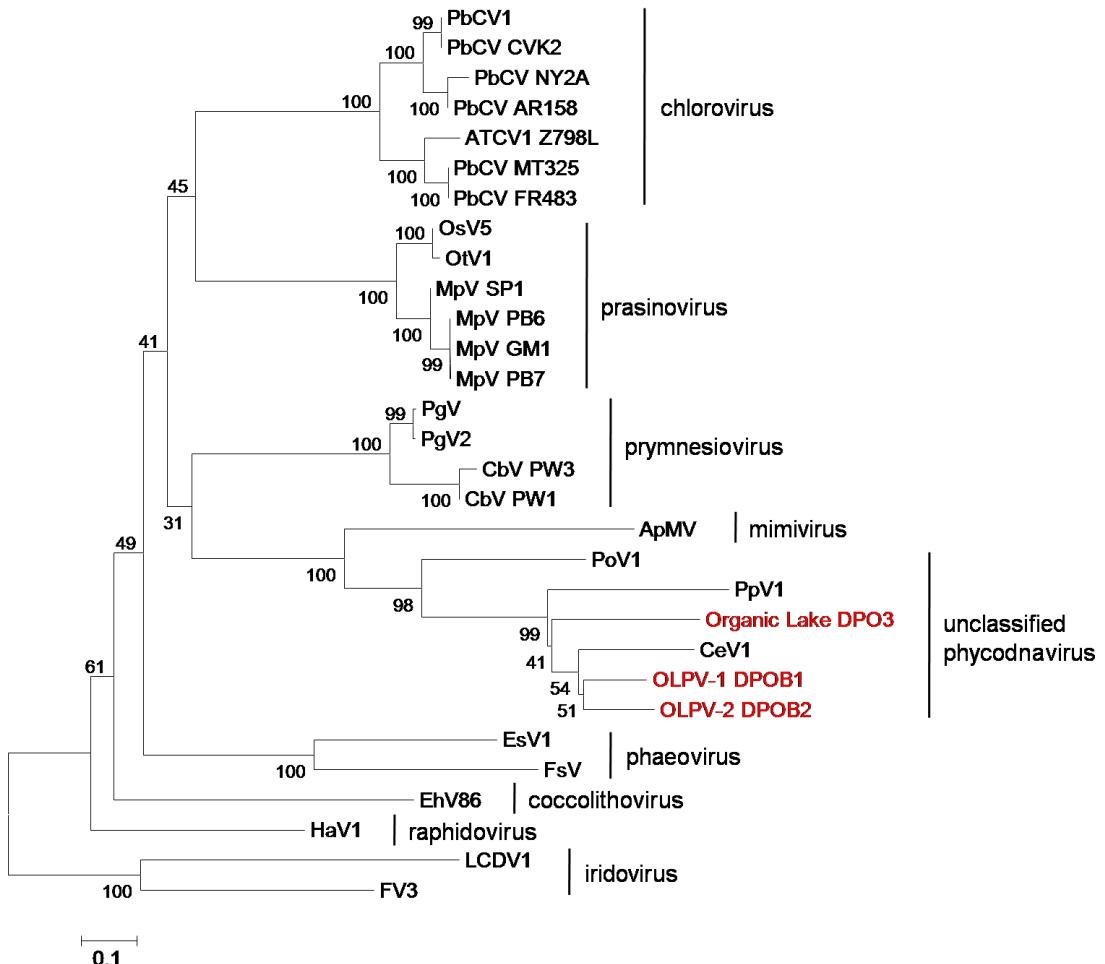


Figure 3.3: Neighbour-joining tree of B family DNA polymerase amino acid sequences from OLPV and NCLDV sequences from GenBank. Organic Lake sequences are shown in red. Abbreviations and accession numbers from bottom to top: PbCV1, *Paramecium bursaria* chlorella virus 1 (AAC00532.1); PbCV CVK2, *P. bursaria* chlorella virus CVK2 (BAA35142.1); PbCV NY2A, *P. bursaria* chlorella virus NY2A (ABT14648.1); PbCV AR158, *P. bursaria* chlorella virus AR158 (ABU43776.1); AtCV1, *Acathocystis turfacea* chlorella virus (ABT16932.1); PbCV MT325, *P. bursaria* chlorella virus MT325(ABT13573.1); PbCV FR483, *P. bursaria* chlorella virus FR483 (ABT15308.1); OsV5, *Ostreococcus* virus 5 (ABY28020.1); OtV1, *O. tauri* virus 1(YP_003495047.1); MpV SP1, *Micromonas pusilla* virus SP1(AAB66713.1); MpV PB6, *M. pusilla* virus PB6 (AAB49743.1); MpV GM1, *M. pusilla* virus GM1 (AAB49742.1); MpV PB7, *M. pusilla* virus PB7 (AAB49744.1); CbV PW1, *Chrysochromulina brevifilum* virus PW1 (AAB49739.1); CbV PW3, *C. brevifilum* virus PW3 (AAB49740.1); ApMV, *Acathamoeba polyphaga* mimivirus (AAV50591.1); PoV, *Pyramimonas orientalis* virus (ABU23717.1); PpV, *Phaeocystis pouchetii* virus (ABU23718.1); CeV1, *C. ericinia* virus 1 (ABU23716.1); EsV1, *Ectocarpus siliculosus* virus (AAK14511.1); FsV, *Feldmannia* sp. virus (AAB67116.1); EhV86, *Emiliania huxleyii* virus 86 (CAI65453.1); HaV1, *Heterosigma akashiwo* virus 1 (BAE06251.1); FV3, Frog virus 3(AAT09720.1) and LCDV1, Lymphocystis disease virus 1(NP_078724.1).

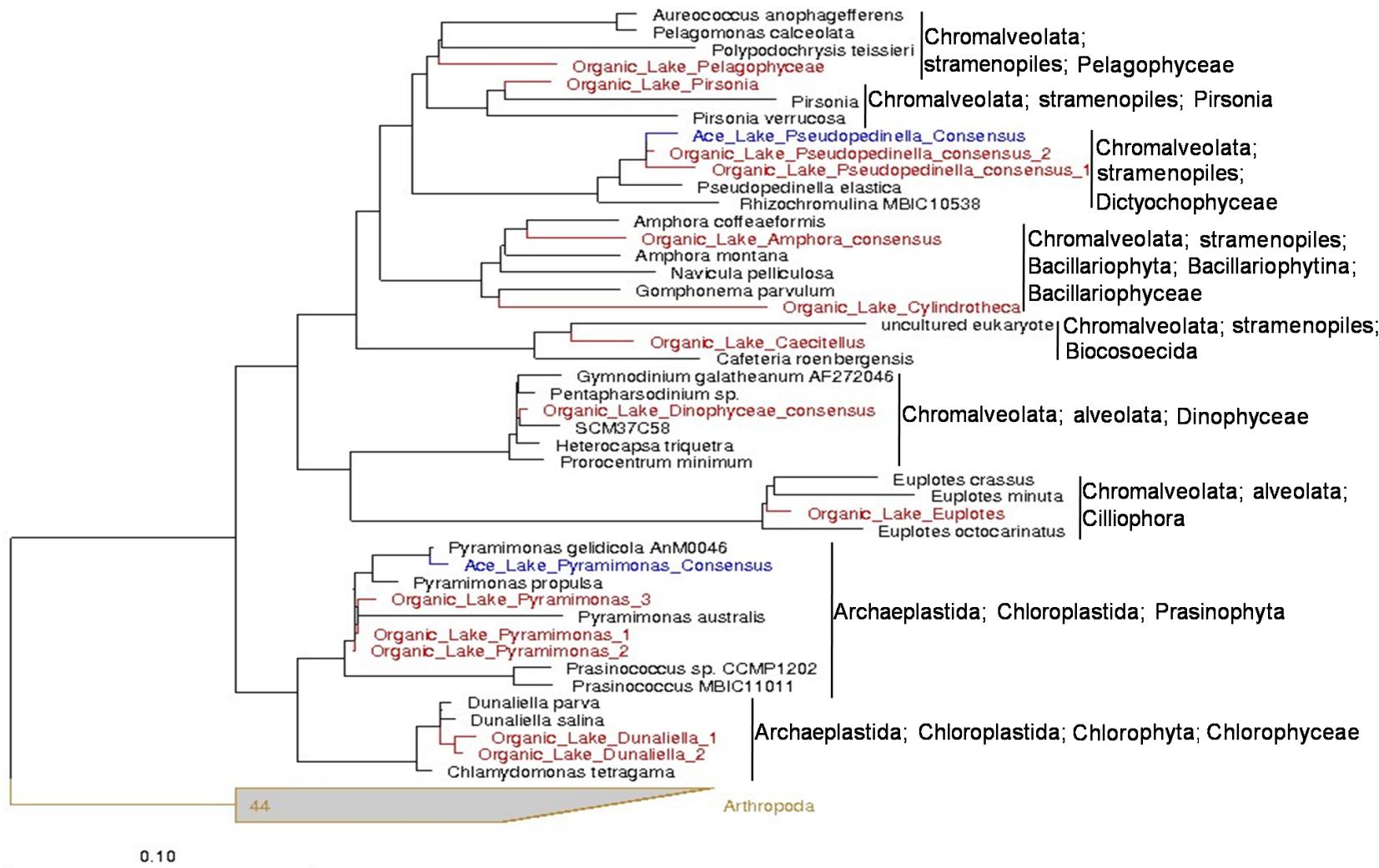


Figure 3.4: Phylogeny of the 18S rRNA genes from the 2006 Organic Lake metagenome. Organic Lake sequences are shown in red. Ace Lake sequences are shown in blue.

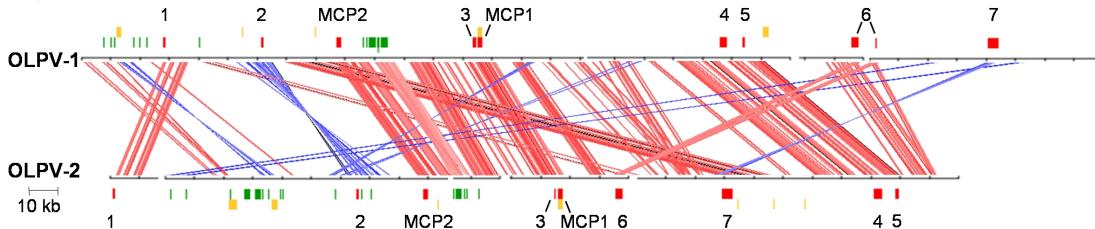


Figure 3.5: Maps of OLPV-1 and OLPV-2 scaffolds and comparison of the location of genes. Genes are marked as follows: single-copy conserved orthologues and MCP (red), regions with identity to OLV (green), proteins identified in the metaproteome (yellow), ribosomal nucleotide reductase β (1), VV A32 packaging ATPase (2), VV VLTF3 transcription factor (3), VV D5 replicative helicase (4), PbCV-1 A482R-like putative transcription factor (5), ribonucleotide reductase α (6), and DNA polymerase B (7). Lines connect homologous regions between the OLPV-1 and OLPV-2 scaffolds in the same orientation (red) and reverse orientation (blue).

Supporting the presence of more than one PV, pairs of single-copy PV orthologues (ribonucleotide reductase α and β subunits, VV A32R virion packaging helicase, PbCV1 A482R-like putative transcription factor, VV D5 ATPase and VLTF2 family transcription factor) were identified in the high coverage scaffolds that shared an average of 81% percent amino acid identity. Based on the positions of single copy genes on the scaffolds and the percent identity between them, the high coverage scaffolds were grouped into two strains designated OLPV-1 and OLPV-2 according to their DPOB phylogeny (Figure 3.3). The remaining high coverage scaffolds were assigned to either strain, resulting in two near-complete genomes of \sim 300 kbp each (Figure 3.5), that are within the range of other sequenced PV genomes (155–407 kbp). In addition, several OLPV genomic fragments contained PV homologues in high coverage scaffolds that could not be confidently assigned to either strain.

Both OLPV strains contain a PpV-like MCP designated MCP1 and another unique MCP designated MCP2 (Figure 3.6). Both OLPV MCP1s were identified in the metaproteome (Figure 3.5 and Table 3.3) but MCP2 was not. In addition to MCPs, the metaproteome contained a range of abundant structural proteins and others more likely to be packaged in the virion (e.g. chaperone), that were expressed by OLPV-1, OLPV-2 and/or an OLPV genomic fragment Table 3.3. These data suggest that MCP1 is the major structural protein, and that both OLPV-1 and OLPV-2 were in a productive cycle in the lake at the time of sampling.

3.4.2 Complete genome of an Organic Lake virophage

Sputnik is a small (50 nm) icosahedral satellite virus of mamavirus (a new strain of ApMV). It was termed a virophage because co-infection with Sputnik is deleterious to the mamavirus, resulting in abnormal virions and a decrease in mamavirus infectivity (25). One 28 kbp scaffold in the low GC high coverage group had six out of 38 predicted proteins homologous to Sputnik virophage proteins (Figure 3.7 and Table 3.4), and one PV homologue. The scaffold had a low GC content (\sim 30%), similar to the Sputnik

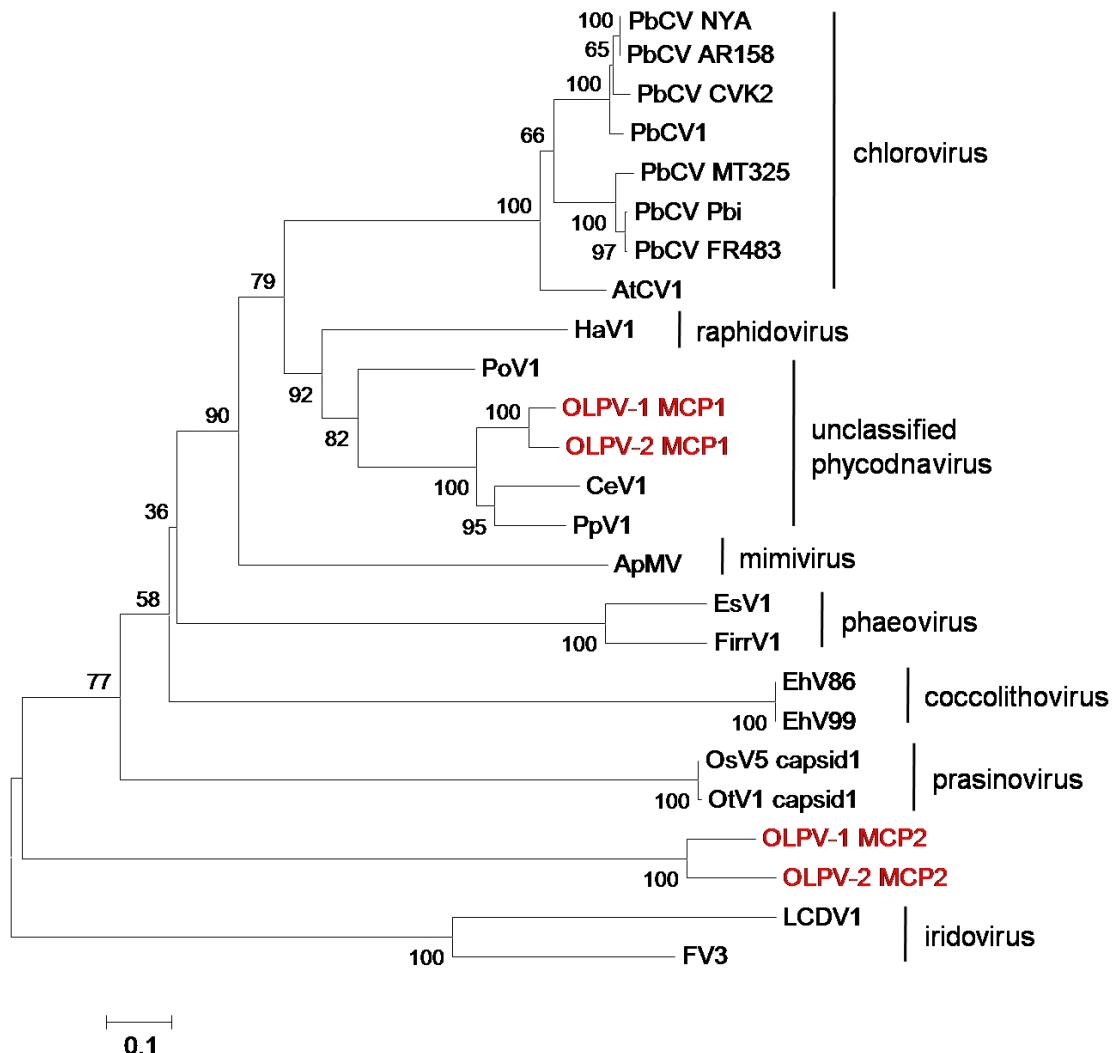


Figure 3.6: Neighbour-joining tree of major capsid protein amino acid sequences from OLPV and other NCLDV sequences from GenBank. Abbreviations and accession numbers from top to bottom: PbCV NYA, *Paramecium bursaria* chlorella virus NY2A (ABT14984.1); PbCV AR158, *P. bursaria* chlorella virus AR158 (ABU44077.1); PbCV CVK2, *P. bursaria* chlorella virus CVK2 (BAA35143.1); PbCV1, *P. bursaria* chlorella virus 1(AAA88828.1); PbCV MT325, *P. bursaria* chlorella virus MT325 (ABT14017.1); PbCV Pbi, *P. bursaria* chlorella virus Pbi (AAC27492.1); PbCV FR483, *P. bursaria* chlorella virus FR483 (ABT15755.1); AtCV1, *Acathocystis turfacea* chlorella virus 1 (ABT16414.1); HaV1, *Heterosigma akashiwo* virus 1 (BAE06835.1); PoV, *Pyramimonas orientalis* virus (ABU23714.1); CeV1, *Chrysochromulina ericinia* virus 1 (ABU23712.1); PpV1, *Phaeocystis pouchetii* virus (ABU23715.1); ApMV, *Acathamoeba polyphaga* virus (Q5UQL7.2); EsV1, *Ectocarpus siliculosus* virus (AAK14534.1); FirrV1, *Feldmannia irregularis* virus 1 (AAR26925.1); EhV86, *E. huxleyi* virus 86 (CAI65508.2); EhV99, *E. huxleyi* virus 99 (ABU23713.1); OsV5, *Ostreococcus* virus 5 (ABY27849.1); OtV1, *O. tauri* virus 1 (CAY39653.1); LCDV1, Lymphocystis diseases virus 1(AAC24486.2); FV3, Frog virus 3 (AAT09750.1).

Table 3.3: OLPV and OLV proteins identified in the December 2006 0.1 µm size fraction metaproteome. Peptide sequences by which the proteins were identified are shown in Appendix Table 6.1. NSA, normalised spectral abundance; Cov., coverage. ^aProteins that have some shared peptides; ^b162322406 and 162276024 are protein homologues; ^cA group of proteins containing similar peptides that could not be differentiated by the mass spectral analysis. Only one gene number of that groups is displayed.

Gene ID	Source	NSA	Accession	Description	Cov. (%)	Peptides (unique)
162322530 ^a	OLPV-1	0.000661	A7U6F0	Major capsid protein [<i>Phaeocystis pouchetii</i> virus]	33	15 (4)
162322348	OLPV-1	0.000120	-	-	11.3	2 (2)
162322406 ^b	OLPV-1	0.000177	-	-	29.4	4 (4)
162313481	OLPV-1	0.000010	YP_002714448	Leucine rich repeat-containing Miro-like protein [<i>Synechococcus</i> sp. PCC7335]	3.96	2 (2)
162276060	OLPV-2	0.000897	-	-	28.9	2 (2)
162300260	OLPV-2	0.000226	-	-	34.6	2 (2)
162276024 ^b	OLPV-2	0.000127	-	-	16	3 (3)
162275992	OLPV-2	0.000098	NP_048709	Hypothetical protein PBCV1_A352L [<i>Paramecium bursaria</i> chlorella virus 1]	16.6	2 (2)
162300108	OLPV-2	0.000046	ZP_01471812	BNR containing hypothetical protein RS9916_28494 [<i>Synechococcus</i> sp. RS9916]	7.66	5 (5)
162319393 ^a	OLPV-2	0.000016	A7U6F0	Major capsid protein [<i>Phaeocystis pouchetii</i> virus]	26.3	13 (2)
162300134 ^c	OLPV-1/2	0.00010	AAR21578	Heat shock protein 70	6.97	3 (3)
162286324 ^c	OLPV	0.000176	NP_048575	[<i>Phytophthora nicotiana</i> e] Hypothetical protein PBCV1_A227L [<i>Paramecium bursaria</i> chlorella virus 1]	14.7	2 (2)
OLV9	OLV	0.001681	YP_002122381	Capsid protein V20 [Sputnik virophage]	31.1	15 (15)
OLV8	OLV	0.000334	YP_002122379	N-term: hypothetical protein V18 [Sputnik virophage]	19.1	8 (8)
			YP_002122380	C-term: minor capsid protein V19 [Sputnik virophage]		

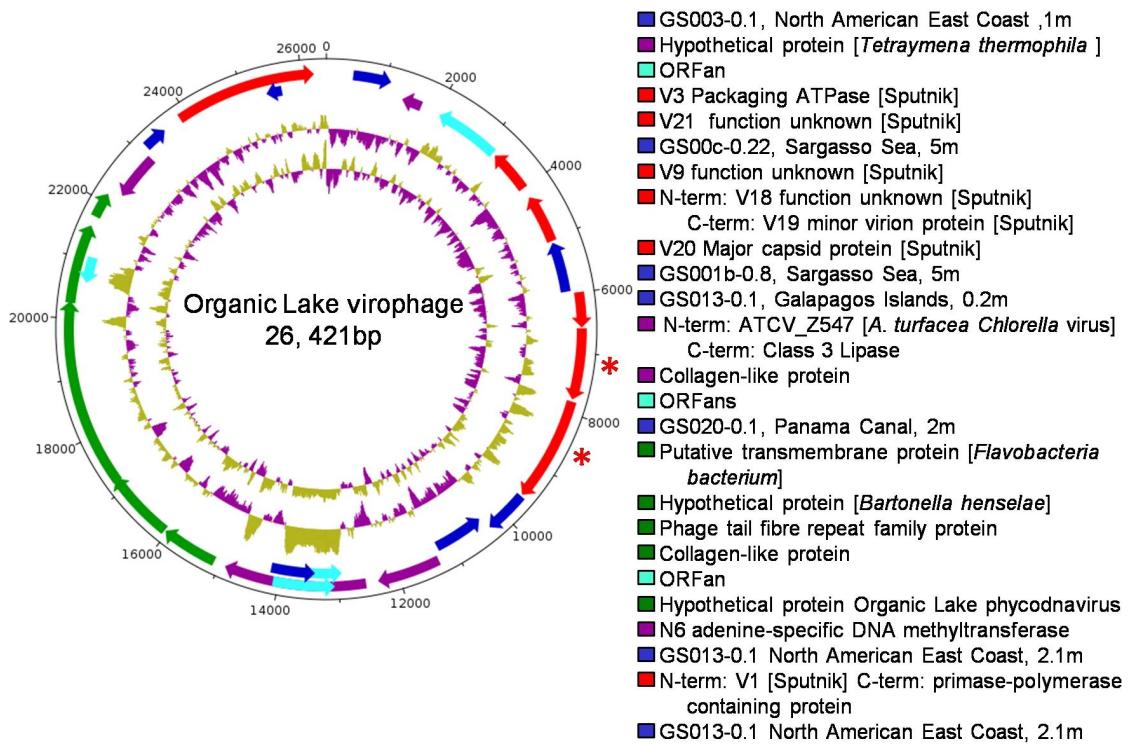


Figure 3.7: Genomic map of Organic Lake virophage. From the outside-in, circles represent, 1) predicted coding sequences on the forward strand, 2) predicted coding sequences on the reverse strand, 3) GC skew, and 4) GC plot. Predicted coding sequences are coloured: Sputnik homologues (red), OLPV homologues (green), non-Sputnik NR homologues (purple), GOS peptide database homologues (blue), and ORFan (cyan). Sequences identified in the metaproteome are marked with an asterisk. Descriptions of the predicted coding sequences from both strands are shown clockwise from position zero.

genome, and was larger in size (28 kbp vs 18 kbp for Sputnik). Using polymerase chain reaction (PCR) and sequencing, the scaffold was found to represent a complete circular virophage genome. This shows the Organic Lake genome has the same circular topology as the Sputnik genome. Virus-like particles resembling Sputnik in size and morphology were identified by TEM (Figure 3.2B).

Table 3.4: Top BLASTP matches of predicted coding sequences from the OLV genome compared to OLPV, NR protein database, and CAMERA metagenomic reads ORF peptide database.

Gene ID	Start	End	NR (acc, %ID, e-value)	OLPV (geneID, %ID, e-value)	CAMERA (acc, %ID, e-value)
OLV1	460	1,077	-	-	GS003, 0.1, North America East Coast, 1 m (JCVI_PEP_1105157870626/41%/1e-29)
OLV2	1,701	1,333	Hypothetical protein [<i>Tetrahymena thermophila</i> SB210] (XP_001029204.1/38% /3e-04)	-	GS012, 0.1, North American East Coast, 13.2 m (JCVI_PEP_1105080106223/42%/1.7e-11)
OLV3	3,187	2,030	-	-	-
OLV4	3,991	3,224	-	-	-
OLV5	5,029	4,160	V3 [Sputnik virophage] (YP_002122364.1/ 39%/4e-24)	-	GS001b, 0.8, Sargasso Sea, 5 m (JCVI_PEP_1105131296011/43%/5e-38)
OLV6	5,940	5,044	-	-	GS000c, 0.22, Sargasso Sea, 5m (JCVI_PEP_1105136847382/24%/6e-3)
42	OLV7	5,978	6,547	V9 [Sputnik virophage] (YP_002122370.1/35%/3e-14)	-
	OLV8	6,574	7,740	N-term: V18 [Sputnik virophage] (YP_002122379.1/27%/9e-05) C-term: V19 [Sputnik virophage] (YP_002122380.1/26%/0.16)	-
OLV9	7,791	9,518	V20 [Sputnik virophage] (YP_002122381.1/28%/9e-10)	-	GS033, 0.1, Galapagos Islands, 0.2 m (JCVI_PEP_1105120114513/28%/2e-14)
OLV10	9,563	10,273	-	-	GS001b, 0.8, Sargasso Sea, 5 m (JCVI_PEP_1105163928413/61%/5e-4)
OLV11	11,210	10,317	-	-	GS013, 0.1, North America East Coast, 2.1 m (JCVI_PEP_1105123792445/39%/9e-37)
OLV12	11,284	12,324	N-term: Hypothetical protein ATCV_Z547R [<i>Acanthocystis turbacea</i> chlorella virus 1]	-	GS018, Caribbean Sea, 1.7 m (JCVI_PEP_1105087988121/34%/5.6e-23)

Continued on next page

Table 3.4 – *Continued from previous page*

Gene ID	Start	End	NR (acc, %ID, e-value)	OLPV (geneID, %ID, e-value)	CAMERA (acc, %ID, e-value)	
			(YP_001427028.1/36%/7e-09) C-term: Lipase class 3 [<i>Bacillus thuringiensis</i> IBL200] (EEM96541.1/27%/1.7e-02)	-	-	
OLV13	12,539	14,884	Collagen-like protein [<i>Bacillus megaterium</i>] (YP_001569009.1/66.67%/4e-03)	-	GS027, 0.1, Galapagos Islands, 2.2m (JCVI_PEP_1105075498120/43%/6.7e-11)	
OLV14	14,023	12,905	-	-	-	
OLV15	13,041	14,078	-	-	-	
OLV16	15,094	13,372	-	-	GS020, 0.1, Panama Canal, 2 m (JCVI_PEP_1105127133835/36%/8.5e-11)	
43	OLV17	15,094	16,023	Putative transmembrane protein [<i>Flavobacteria</i> bacterium BAL38] (ZP_01734433.1/51%/8e-34)	Lipoprotein Q-like protein (162322444/40%/1e-24)	GS009, 0.1, North American East Coast, 1 m (JCVI_PEP_1105137954859/50%/4e-37)
	OLV18	16,054	17,211	Hypothetical protein BH13620 [<i>Bartonella henselae</i> str. Houston-1] (YP_034083.1/15%/4e-04)	<i>Cyanothece</i> sp. cce_0037-like protein (162322244/65%/2e-33)	GS000c, 0.1, Caribbean Sea, 2 m (JCVI_PEP_1105149563549/39%/2e-26)
	OLV19	17,168	20,278	Phage tail fiber repeat family protein [<i>Trichomonas vaginalis</i> G3] (XP_001296018.1/42%/4e-11)	Lipoprotein Q-like protein (162322444/65%/9e-33)	GS016, 0.1, Caribbean Sea, 2 m (JCVI_PEP_1105149563549/29%/1e-27)
	OLV20	20,266	21,570	Collagen triple helix containing protein A1Q_3499 [<i>Vibrio harveyi</i> HY01] (ZP_01986098.1/69%/6e-04)	Hypothetical protein (162322252/32%/1e-07)	GS033, 0.1, Galapagos Islands, 0.2 m (JCVI_PEP_1105153074955/69%/1e-5)
	OLV21	21,089	20,622	-	-	-
	OLV22	21,747	22,157	-	Hypothetical protein (162322266/56%/5e-31)	GS017, 0.1, Caribbean Sea, 2 m (JCVI_PEP_1105100448171/43%/4e-24)
	OLV23	23,089	22,256	D12 class N6 adenine-specific DNA methyltransferase [" <i>Candidatus Koribacter versatilis</i> " Ellin345] (YP_592471.1/28%/1e-24)	-	GS002, 0.1, North America East Coast, 1 m (JCVI_PEP_1105085453201/33%/8e-18)

Continued on next page

Table 3.4 – *Continued from previous page*

Gene ID	Start	End	NR (acc, %ID, e-value)	OLPV (geneID, %ID, e-value)	CAMERA (acc, %ID, e-value)
OLV24	23,174	23,560	-	-	GS013, 0.1, North American East Coast 2.1 m (JCVI_PEP_1105132174179/32%/1e-03)
OLV25	23,889	26,219	N-term: V13 [Sputnik virophage] (YP_002122374.1/34%/5e-31) C-term: Primase-polymerase domain containing hypothetical protein [<i>Ostreococcus lucimarinus</i> CCE9901] (XP_001421479.1/29%/9e-32)	-	GS030, 0.1, Galapagos Islands, 19 m (JCVI_PEP_1105105378071/40%/8e-38) GS013, 0.1, North America East Coast, 2.1 m (JCVI_PEP_1105129419397/51%/8e-71)
OLV26	25,666	25,376	-	-	GS013, 0.1, North American East Coast, 2.1 m (JCVI_PEP_1105129419399/54%/8e-6)

Sputnik homologues present in the Organic Lake scaffold included the V20 MCP, V3 DNA packaging ATPase, V13 putative DNA polymerase/primase and others of unknown function (V9, V18, V21 and V32) (Figure 3.7 and (Table 3.3)). The OLV is distinct to Sputnik as proteins share 27–42% amino acid identity (28% MCP identity). OLV proteins include OLV9, the homologue of Sputnik V20 MCP, and OLV8, a fusion of the uncharacterised V18 and minor virion protein V19 from Sputnik (Figure 3.7 and (Table 3.4)). The large number of homologues, including genes that fulfill essential functions in Sputnik (V20, V3 and V13), indicate that OLV and Sputnik have physiological similarities.

3.4.3 Gene exchange between virophage and phycodnaviruses

As PVs are related to ApMV (27) and are abundant in Organic Lake, it stands to reason that OLPV is the helper of OLV. In the OLV genome, OLV12 is a chlorella virus-derived gene, indicating that gene exchange has occurred between OLV and PVs (the function of OLV12 is discussed below). Similar observations were made for Sputnik, which carries four genes (V6, V7, V12 and V13) in common with the mamavirus, indicative of gene exchange between the viruses and possible co-evolution (25). As the V6, V7, V12 and V13 proteins have been associated with virophage-helper specificity, we reasoned that the functional analogues in OLV would have highest identity to proteins from its helper virus, rather than Sputnik.

By comparing OLV and OLPV, a 7,408 bp region was identified in OLV encoding five proteins (OLV17–22) with identity (32–65%) to sequences in both OLPV-1 and OLPV-2 (Figure 3.5, Figure 3.8 and Table 3.4). OLV20 and OLV13 are collagen triple-helix-repeat-containing proteins, analogous to Sputnik collagen-like proteins (V6 and V7) involved in protein–protein interactions in the ApMV virus factory. Sputnik can replicate with either mamavirus or ApMV as a helper, although coinfection rates are higher with the mamavirus. V6 is the only protein with higher identity (69%) to mamavirus than ApMV (42%) (25). Since OLV20 has equivalent identity (63%) with OLPV-1 and OLPV-2, it appears that OLV may be capable of interacting with both OLPV strains. Also within the conserved region, OLV22, is a 141 aa protein of unknown function that only matches sequences from OLPV and the GOS expedition (Table 3.4). Similar to OLV22, Sputnik V12 is a small protein (152 aa) of unknown function with high identity to ApMV, and both may mediate a specific helper–virophage interaction. Other genes in this region of OLV can be mapped to OLPV, including a putative transmembrane protein (OLV17) and paralogous phage tail fibre repeat containing proteins, OLV18 and OLV19. Analogous to the collagen-like proteins, OLV19 and OLV20 probably facilitate interactions between helper and virophage.

OLV12, which is unique to OLV, consists of a C-terminal domain present in conserved hypothetical chlorella virus proteins and an N-terminal domain most closely related to class 3 lipases that may confer OLV selectivity to a PV. OLV12 may function similarly to the Sputnik V15 membrane protein in modifying the ApMV membrane (25). The Sputnik V13 consists of a primase domain and SF3 helicase domain related to NCLDV

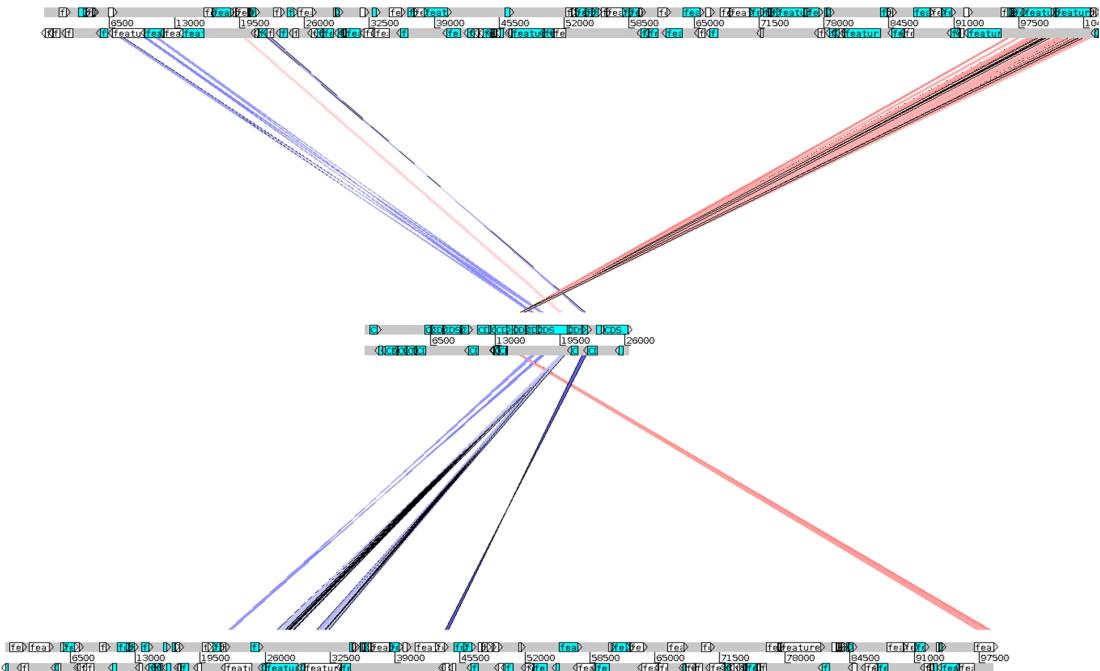


Figure 3.8: Comparison of the location of genes in OLV compared to OLPVs. OLV (centre), OLPV-1 (top), and **OVLP!-2** (bottom).

homologues, involved in DNA replication. The helicase domain of OLV25 and V13 are similar, although the primase domain is more similar to a protein from *Ostreococcus lucimarinus*, implying a past association of OLV with a prasinophyte alga host.

Genes unique to OLV point to adaptations specific to its helper–host system. Most notably, OLV possesses a N6 adenine-specific DNA methyltransferase, as does OLPV. In OLPV-1, genes for a bacterial type I restriction modification (RM) system are adjacent to a gene encoding a type I methylase-S target recognition domain protein, and upstream of a DNA helicase distantly related to type III restriction endonuclease (RE) subunits. A large number of chlorella virus genomes have both 5mC and 6mA methylation (33), and several contain functional RM systems (34). The prototype chlorella virus PbCV-1 possesses REs packaged in the virion for degrading host DNA soon after infection (35). In contrast to OLV and OLPV, DNA methyltransferases are absent in both Sputnik and ApMV, indicating that the N6 adenine-specific DNA methyltransferase has been selected in OLV to reduce endonucleolytic attack mediated by OLPV.

3.4.4 Role of virophage in algal host–phycodnavirus dynamics

The presence of the virophage adds an additional consideration to the microbial loop dynamics. In batch amoeba cultures, co-infection of amoeba with ApMV and Sputnik causes a 70% decrease in infective ApMV particles and a 3-fold decrease in lysis (25). To test how OLV affects OLPV and host population dynamics, we modelled the OLV as an additional predator of a predator in a Lotka–Volterra simulation (Figure 3.9).

The classic Lotka–Volterra model (?) is based on a pair of first-order, non-linear,

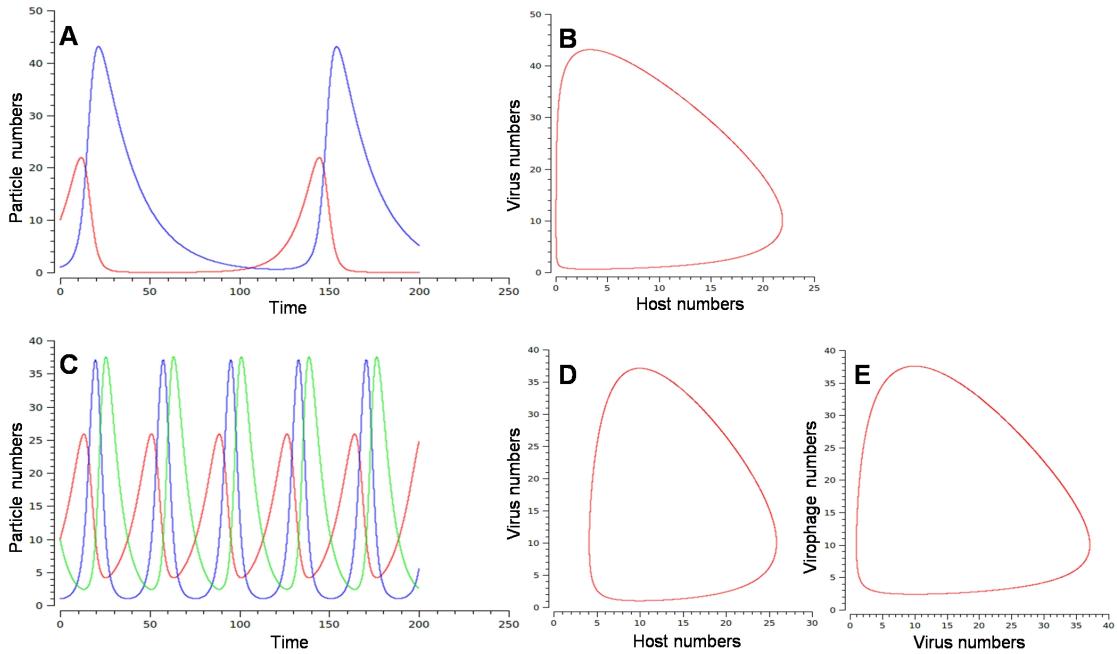


Figure 3.9: Extended Lotka-Volterra models of host–OLPV–OLV population dynamics. **(A)** Time course of host (red line) and OLPV (blue line) populations in the absence of OLV. **(B)** Orbit plot between host and OLPV populations in the absence of OLV with the host and OLPV populations approaching zero during an equilibrium cycle. **(C)** Time course describing the effect of the addition of OLV (green line) on OLPV–host population dynamics as a predator of predator resulting in increased frequency of population oscillations and a higher minimal number of hosts and OLPVs **(D)** compared to in the absence of OLV **(B)**. **(E)** The orbit plot of OLPV and OLV is also shown. Note that the time intervals are arbitrary.

differential equations that can be used to describe the periodic oscillation of the populations of a predator and its prey (?). An example of how predator (virus) populations follows that of its prey (host) populations over time is shown in Figure 3.9A where the populations are at equilibrium. The extended model shown (Figure 3.9C) is based on three equations describing the host (prey), virus (predator) and virophage (predator of predator) interactions. In this model, the effect of virophage is robust, with equilibrium solutions across a wide range of parameter values (Figure 3.9C shows one equilibrium solution). It shows the virus population following that of its host and the virophage population in turn following that of its helper virus or “host”. While the absolute number of hosts do not increase greatly as a result of OLV preying on OLPV, the frequency of host blooms increases in the presence of OLV.

This is due to OLV decreasing the number of infective OLPVs, thereby shortening the recovery time of the host population (Figure 3.9C). This is evident in the orbit plot (Figure 3.9D) as the shift of the orbit away from the axis.

The model reveals that the virophage stimulates the flux of secondary production through the microbial loop by reducing overall mortality of the host algal cell following a bloom, and by increasing the frequency of blooms during the summer light periods. Antarctic lake systems have evolved mechanisms to cope with long light-dark cycles

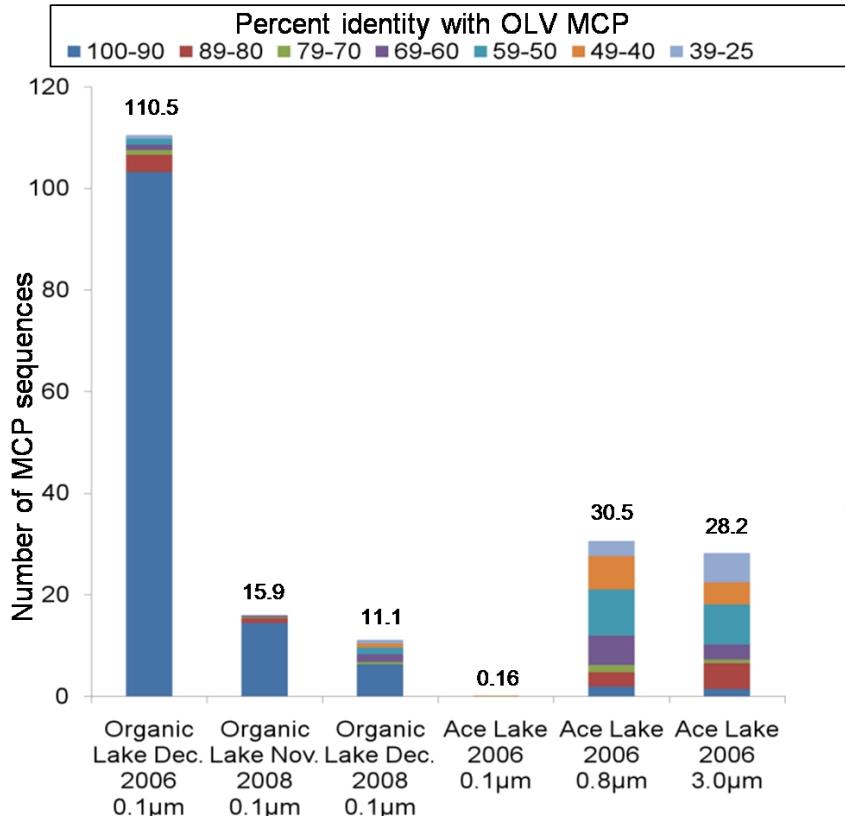


Figure 3.10: Abundance and diversity of virophage capsid proteins in environmental samples. Number of ORFs from metagenomic reads that match to OLV MCP (BLASTP e-value cut-off 1e⁻⁵, abundance normalised to 100,000 reads per sample), and the proportion of virophage capsid types, for the Organic Lake 0.1 μm and Ace Lake 0.1, 0.8 and 3.0 μm fractions.

(14) and shortened trophic chains. In Organic Lake and similar systems, a decrease in PV virulence may be instrumental in maintaining stability of the microbial food web. In other words, the increased turnover in the microbial loop during the extended light periods of the polar summer may help to maintain microbial populations in the lake.

3.4.5 Ecological relevance of virophages in aquatic systems

Metagenomic analysis of Organic Lake samples taken two years later in November (when the lake was ice covered) and December 2008 (partially ice-free) revealed sequences with 99% amino acid identity to OLV MCP indicating persistence of OLV in the ecosystem (Figure 3.10 and Table 3.5). In addition, sequences with lower identity (25–90%) were detected, particularly in December, demonstrating Organic Lake virophages are highly diverse but OLV remained the dominant type.

From surface water samples of nearby Ace Lake (meromictic, surface 2% salinity), a large number of sequences were obtained that matched both the OLV MCP (Figure 3.10, Figure 3.11 and Table 3.5) and PVs (14). All Ace Lake size fractions contained matches to OLV MCP, some with high identity (80–100%) and the majority with greater variation (25–80% identity) (Figure 3.11 and Table 3.5). In contrast to Organic Lake where the largest number of matches was to the 0.1 μm size fraction, the majority of Ace Lake

Table 3.5: BLASTP matches for OLV MCP in predicted ORFs of Organic Lake and Ace Lake contigs and CAMERA metagenomic read ORF peptide database. (E-value cut-off $1e-5$, alignment length $>100\text{aa}$). Aln., alignment length; Cov., coverage.

Sample	Size (μm)	Gene ID	Scaffold ID	Id. (%)	Aln. (aa)	E- value	Cov. (×)
Organic Lake	0.1	OLV9	-	-	-	-	77.12
December	0.8	176157210	scf7180000034275	98.61	575	0.0	16.03
2006	3.0	181703798	deg7180000108904	98.96	575	0.0	48.65
Organic Lake	0.1	192841413	deg7180000116398	99.64	555	0.0	16.03
November	0.8	193037024	scf7180000086663	93.98	133	2e-61	2.5
2008	3.0	192638971	deg7180000028400	99.36	156	1e-76	1.86
		192955191	deg7180000024244	93.98	133	1e-61	3.10
Organic Lake	0.1	192709908	scf7180000109753	99.01	304	9e-173	4.38
December		192709920	scf7180000109753	99.59	244	1e-120	4.38
2008		192712009	deg7180000067104	54.70	117	3e-30	1.58
		192890551	deg7180000061276	36.89	122	2e-13	3.15
	0.8	193060302	deg7180000053149	53.75	160	6e-43	2.30
Ace Lake	0.1	167813925	scf7180000126822	28.86	246	3e-14	2.36
2006		167858124	scf7180000129064	21.78	381	5e-10	1.94
		167891594	scf7180000136823	24.85	326	2e-04	8.15
		167875536	deg7180000086604	22.95	244	6e-04	2.21
	0.8	176091445	deg7180000053588	91.61	143	8e-78	3.35
		175769103	deg7180000078701	88.24	153	1e-74	1.77
		176042318	deg7180000058177	81.77	181	1e-74	2.48
		176000635	deg7180000087166	53.39	221	8e-58	2.50
		176042707	deg7180000058207	50.78	193	5e-46	2.34
		175886340	deg7180000074162	58.90	146	4e-45	2.73
		176249679	deg7180000049481	61.94	155	2e-44	2.75
		175748439	deg7180000058552	76.79	112	2e-35	2.39
		175637390	deg7180000058712	50.91	165	2e-35	2.03
		176100822	deg7180000055966	53.38	133	6e-35	1.48
		176018109	deg7180000086684	59.68	124	4e-27	1.66
		176000624	deg7180000087165	53.85	104	6e-27	1.73
		175805608	deg7180000054222	48.60	107	7e-21	1.93
		175908895	deg7180000061971	51.91	131	4e-20	3.27
		175821062	deg7180000080443	46.46	127	6e-20	1.43
		176026419	deg7180000054364	52.59	116	8e-19	1.47
		176133336	scf7180000089989	38.36	146	3e-12	1.51
		176018257	deg7180000086719	31.21	173	4e-12	1.23
		176137412	deg7180000052688	29.37	126	1e-06	2.47
		175686880	scf7180000092161	24.00	125	2e-06	2.98
	3.0	175741076	deg7180000030508	85.78	232	8e-109	1.25
		175748837	deg7180000027929	55.29	170	4e-44	1.32
		175751996	deg7180000037324	51.27	158	5e-41	1.63
		175859792	scf7180000045944	30.21	288	8e-26	2.69
Punta Cormorant	0.1	JCVI_PEP_1105120114513	-	27.84	273	2e-14	-
hypersaline		JCVI_PEP_1105100621559	-	24.76	307	9e-10	-
lagoon (GS003)		JCVI_PEP_1105161421335	-	25.61	289	2e-6	-
Delaware Bay	0.1	JCVI_PEP_1105106741177	-	24.62	264	6e-14	-
(GS011)		JCVI_PEP_1105089715877	-	27.16	313	1e-17	-
Upwelling	0.1	JCVI_PEP_1105079267881	-	28.23	170	8e-11	-
(GS031)							
Lake Gatun	0.1	JCVI_PEP_1105119255775	-	26.71	149	5e-9	-
Panama (GS020)							

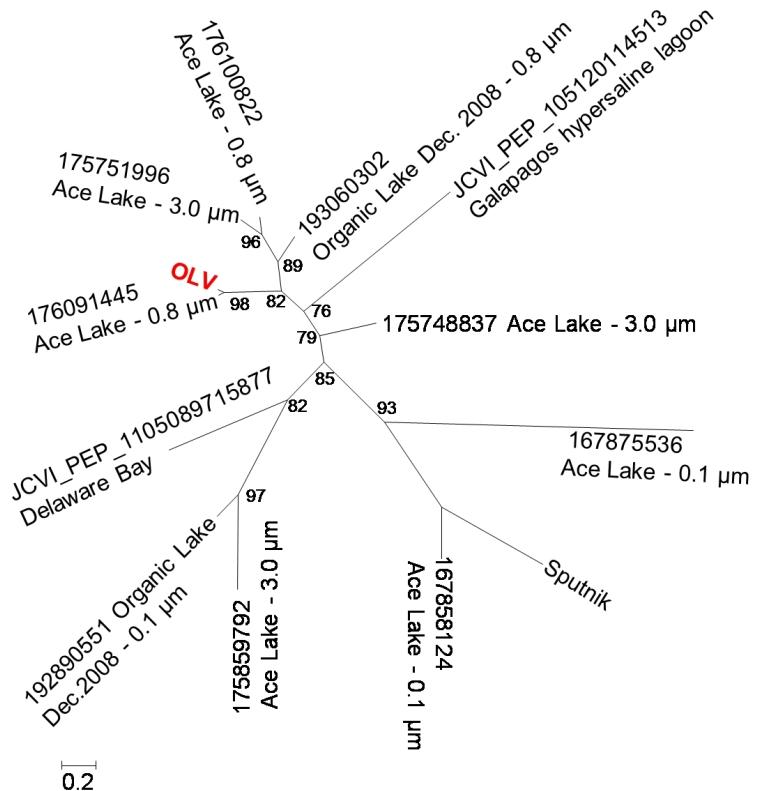


Figure 3.11: Maximum likelihood phylogenetic tree of a conserved 103 aa region of the MCP from Organic Lake, Ace Lake and GOS metagenome data and Sputnik.

sequences were from the larger fractions (Figure 3.10 and Table 3.5). This indicates the Ace Lake virophages were associated with host cells during sampling, or possibly with helper viruses that are larger than the OLPVs.

Extending the OLV MCP search to the GOS data revealed matches (25–28% identity) to sequences from the hypersaline Punta Cormorant Lagoon (Floreana Island, Galapagos), an oceanic upwelling near Fernandina Island (Galapagos), Delaware Bay estuary (NJ, USA), and freshwater Lake Gatun (Panama) (Table 3.5). The phylogenetic analysis of a conserved 103 amino acid region of the MCPs revealed a number of clusters, with Sputnik clustering with virophage sequences from Ace Lake that had low identity (22%) to OLV MCP (Figure 3.11). To improve searches for virophages and better understand their physiology and evolution, it will be valuable to target more genomes (e.g. the Ace Lake 167858124 relative with 40% MCP identity to Sputnik) and determine which genes are core to virophages and what relationship exists between genome complement and MCP identity.

In view of the implications of the virophage modelling (Figure 3.9), the abundance and persistence of OLV in Organic Lake and the presence of diverse virophage signatures in a variety of lake systems (fresh to hypersaline), an estuary, an ocean upwelling site and a water cooling tower (Sputnik), our study indicates that numerous types of virophages exist and play a previously unrecognised role in regulating host–virus interactions and influencing ecosystem function in aquatic environments.

3.5 Acknowledgements

We thank Craig Venter, John Bowman, Louise (Cromer) Newman, Anthony Hull, John Rich and Martin Riddle for providing helpful discussion and logistical support associated with the Antarctic expedition, and Lisa Ziegler for discussion about marine viruses. We acknowledge technical support for computing infrastructure and software development from Intersect, and in particular assistance from Joachim Mai. This work was supported by the Australian Research Council and the Australian Antarctic Division. Funding for sequencing was provided by the Gordon and Betty Moore Foundation to the J. Craig Venter Institute. Mass spectrometric results were obtained at the Bioanalytical Mass Spectrometry Facility within the Analytical Centre of the University of New South Wales. This work was undertaken using infrastructure provided by NSW Government co-investment in the National Collaborative Research Infrastructure Scheme. Subsidized access to this facility is gratefully acknowledged. We thank Jenny Norman from the UNSW Electron Microscopy Unit her assistance in generating images.

Chapter 4

Strategies of carbon conservation and unusual sulphur biogeochemistry in a hypersaline Antarctic Lake

Co-authorship Statement

A version of this chapter has been submitted as:

Sheree Yau, Federico M. Lauro, Timothy J. Williams, Matthew Z. DeMaere, Mark V. Brown, John Rich, John A.E. Gibson, Ricardo Cavicchioli. Strategies of carbon conservation and unusual sulfur biogeochemistry in a hypersaline lake. *ISME Journal* (submitted), 2013.

Contributions to this manuscript by other researchers is as follows.

Research was designed by Federico Lauro, Mark Brown, John Gibson and Ricardo Cavicchioli. Sample collection was performed by Federico Lauro, Mark Brown and Ricardo Cavicchioli. Metagenomic sequence filtering, global assembly and annotation was performed by Matthew DeMaere. Assistance in interpretation of functional potential provided by Timothy Williams.

Apart from these contributions, I performed all other data analyses and interpretations.

4.1 Abstract

Organic Lake is a shallow, marine-derived hypersaline lake in the Vestfold Hills, Antarctica that has the highest reported concentration of dimethylsulphide (DMS) in a natural body of water. To determine the composition and functional potential of the microbial community and learn about the unusual sulphur chemistry in Organic Lake, shotgun metagenomics was performed on size fractionated samples collected along a depth profile. Eucaryal phytoflagellates were the main photosynthetic organisms. Bacteria were dominated by the globally distributed heterotrophic taxa *Marinobacter*, *Roseovarius* and *Psychroflexus*. The dominance of heterotrophic degradation coupled with low fixation potential indicates possible net carbon loss. However, abundant marker genes for aerobic anoxygenic phototrophy, sulphur oxidation, rhodopsins and CO oxidation were also linked to the dominant heterotrophic bacteria and indicate use of photo- and lithoheterotrophy as mechanisms for conserving organic carbon. Similarly, a high genetic potential for the recycling of nitrogen compounds likely functions to retain fixed nitrogen in the lake. dimethylsulphopropionate (DMSP) lyase genes (*dddD*, *dddL* and *ddDP*) were abundant indicating DMSP is a significant carbon and energy source. Unlike marine environments, DMSP demethylases (*dmdA*) were less abundant than DMSP lyases indicating that DMSP cleavage is the likely source of the high DMS concentration. DMSP cleavage, photoheterotrophy, lithoheterotrophy and nitrogen remineralisation by dominant Organic Lake bacteria are potentially important adaptations to nutrient constraints. In particular, photo- and lithoheterotrophy reduces the extent of carbon oxidation for energy production allowing more carbon to be used for biosynthetic processes. The study sheds light on how the microbial community in Organic Lake has adapted to the unique physical and chemical properties of this Antarctic Lake environment.

4.2 Introduction

Due to the polar light cycle, phototrophic growth in Antarctic environments is relatively high in summer and negligible in winter (Laybourn-Parry *et al.*, 2005) and requires microbial life to survive under long periods under a scarcity of resources. To overcome this limitation, Eucaryotic phytoflagellates in Ace Lake engage in carbon mixotrophy by grazing on bacterioplankton to supplement their carbon requirements in the winter (Laybourn-Parry *et al.*, 2005). Marine heterotrophic bacteria are known to be similarly resourceful by exploiting light energy through photoheterotrophy that includes aerobic anoxygenic photosynthesis (AAnP) or via use of rhodopsins, or lithoheterotrophy such as oxidation of carbon monoxide (Moran and Miller, 2007). Heterotrophic bacteria that can harness energy sources apart from organic carbon can direct a greater proportion of carbon towards growth, which serves to conserve fixed carbon within a closed systems (Moran and Miller, 2007).

Organic Lake is shallow (~ 7 m) and has variable surface water temperatures (-14 to $+15^\circ\text{C}$) while remaining sub-zero throughout most of its depth (Franzmann *et al.*, 1987b; Gibson *et al.*, 1991; Roberts *et al.*, 1993; Gibson, 1999). The salt and marine biota in the lake originate from seawater that was trapped in a basin $\sim 3,000$ BP (Zwartz *et al.*, 1998; Bird *et al.*, 1991). The bottom waters of Organic Lake are unusual due to the absence of hydrogen sulphide and the high concentration of the volatile gas dimethylsulphide (DMS) (Deprez *et al.*, 1986; Franzmann *et al.*, 1987b; Gibson *et al.*, 1991; Roberts and Burton, 1993; Roberts *et al.*, 1993). Concentrations of DMS as high as 5,000 nM have been recorded in Organic Lake (Gibson *et al.*, 1991), 100 times the maximum concentration recorded from seawater in the adjacent Prydz Bay and at least 1,000 times that of the open Southern Ocean (Curran *et al.*, 1998). More than forty years ago atmospheric DMS was proposed to have a regulatory effect on global cloud cover as it is a precursor of cloud condensation nuclei (Lovelock and Maggs, 1972; Charlson *et al.*, 1987). However, the first enzymes involved in DMS production were only identified in the last six years (Todd *et al.*, 2007). Rapid progress has been made in this short period and the pathways and organisms involved in DMS transformations have been extensively reviewed (Johnston *et al.*, 2008; Schäfer *et al.*, 2010; Curson *et al.*, 2011; Reisch *et al.*, 2011; Moran *et al.*, 2012).

The main source of DMS in the marine environment is from the breakdown of dimethylsulphopropionate (DMSP). Eucaryal phytoplankton, in particular diatoms, dinoflagellates and haptophytes, produce large quantities of DMSP, an organo-sulphur compound that is thought to function principally as an osmolyte. DMSP is released due to cell lysis, grazing or leakage and follows two known fates: DMSP cleavage by DMSP lyases (DddD, -L, -P, -Q, -W and -Y) or demethylation by DMSP demethylase (DmdA). Both pathways are associated with diverse microorganisms that can utilize DMSP as a sole carbon and energy source. However, it is only the cleavage pathway that releases volatile DMS that can lead to sulphur loss through ventilation to the atmosphere. The very high levels of DMS in Organic Lake make it an ideal system for identifying the microorganisms and the processes involved in DMS accumulation.

The previous Organic Lake metagenomic study examined viruses from the 0.1 µm fraction of surface water that was collected from Organic Lake in December 2006, and November and December 2008 (Yau *et al.*, 2011) (Chapter 3). In the present study we focused on the cellular population rather than viruses. Our study determined the composition and functional potential of Organic Lake microbiota and, in conjunction with historic and contemporary physico-chemical data, generated an integrative understanding of the whole lake ecosystem.

4.3 Materials and methods

4.3.1 Characteristics of the lake and sample collection

The water level of Organic Lake was measured by surveying as +1.886 m relative to the survey mark (NMV / S / 53) located at 68°27'23.4"S, 78°11'22.6"E. Water was collected from Organic Lake on 10 November 2008 through a 30 cm hole in the 0.8 m thick ice cover above the deepest point in the lake. The sampling hole was established at 68°27'22.2"S, 78°11'23.9"E following bathymetry measurements constructed on a metric grid. Samples were collected for metagenomics, microscopy and chemical analyses at 1.7, 4.2, 5.7, 6.5 and 6.7 m depths (maximum lake depth 6.8 m).

For metagenomics, lake water was passed through a 20 µm pore size pre-filter, and microbial biomass captured by sequential filtration onto 3.0 µm, 0.8 µm and 0.1 µm pore size 293 mm polyethersulfone membrane filters, and samples immediately preserved in buffer and cryogenically frozen in liquid nitrogen, as described previously (Ng *et al.*, 2010; Lauro *et al.*, 2011). Between 1–2 L of lake water was sufficient to saturate the holding capacity of the filters. DNA was extracted from the filters, samples sequenced using the Roche GS-FLX titanium sequencer, and reads processed to remove low quality bases, assembled and annotated, as previously described (Ng *et al.*, 2010; Lauro *et al.*, 2011). A summary of the 2.4 Gbp of metagenomic data is provided in Table 4.1.

4.3.2 Physical and chemical analyses

An *in situ* profile of pH, conductivity, turbidity, dissolved oxygen (DO) and pressure was measured using a submersible probe (YSI sonde model V6600). A temperature profile was measured using a maximum-minimum mercury thermometer as the YSI probe did not have a capacity to record temperature below –10°C. The 5.7 m sample corresponded to the turbidity maximum and the 6.5 m sample to the turbidity minimum. Conductivity at *in situ* temperature was converted to conductivity at 15°C as described previously (Gibson, 1999). The adjusted conductivity brings the temperature to within a range suitable for estimating practical salinity using the formula of Fofonoff and Millard (1983). Salinity was likely to have been underestimated as it is higher than the range (2–42) for which the conductivity–salinity relation holds. However, the relative difference in salinity between the samples would be accurate.

Density was calculated from the *in situ* conductivity and temperature using the equations described by Gibson *et al.* (1990) and expressed at temperature T as:

Table 4.1: Summary of metagenomic data for Organic Lake November 2008 profile.

ID	Depth (m)	Size (μm)	Trimmed reads	Predicted ORFs (%KEGG Scaffolds matches)	Scaffolds (reads)	>10 kbp scaffolds (reads)	Annotated scaffold ORFs (total ORFs)
GS374	1.7	0.1	494,573	533,468 (31)	4,318 (63,194)	5 (771)	33,262 (83,684)
		0.8	472,635	470,949 (52)	4,161 (126,519)	68 (17,061)	37,857 (63,140)
		3.0	158,121	158,573 (50)	2,584 (39,591)	4 (520)	18,126 (28,425)
GS375	4.2	0.1	541,962	556,791 (30)	4,899 (80,316)	2 (232)	35,318 (87,631)
		0.8	472,570	492,130 (53)	5,104 (127,243)	80 (18,461)	42,508 (68,366)
		3.0	321,112	324,365 (56)	3,983 (98,102)	69 (14,713)	30,938 (51,452)
GS376	5.7	0.1	363,280	387,528 (25)	2,342 (39,422)	6 (1,801)	21,798 (61,595)
		0.8	484,635	448,373 (59)	6,820 (152,646)	134 (29,903)	47,846 (73, 282)
		3.0	290,428	292,358 (51)	3,571 (77,277)	58 (10,231)	28,199 (48,910)
GS377	6.5	0.1	497,363	572,892 (29)	5,029 (80,520)	14 (2,711)	36,685 (92,420)
		0.8	465,381	454,018 (51)	4,202 (129,193)	57 (17,004)	43,852 (70,382)
		3.0	187,045	211,354 (59)	2,100 (60,636)	51 (9,321)	20,713 (33,497)
GS378	6.7	0.1	516,870	586,375 (26)	3,694 (58,618)	14 (3,422)	33,243 (96,334)
		0.8	548,253	626,115 (57)	6,957 (161,202)	136 (32,889)	56,452 (88,738)
		3.0	202,310	219,992 (58)	2,304 (66,389)	57 (11,167)	22,786 (35,034)

$$\sigma_T = (1000 - \text{density}) \text{ kg m}^{-3}$$

Ammonia, nitrate, nitrite, total nitrogen (TN), total dissolved nitrogen (TDN), dissolved reactive phosphorus (DRP), total phosphorus (TP), total dissolved phosphorus (TDP), total organic carbon (TOC), dissolved organic carbon (DOC), total sulphur (TS) and total dissolved sulphur (TDS) were determined by American Public Health Associations Standard Methods at the Analytical Services, Tasmania. Values for dissolved nutrients were measured after filtration through a 0.1 μm pore size membrane filter. All other nutrients were measured from water collected after filtration through the on-site 20 μm pore size pre-filter.

Ammonia, nitrate, nitrite, DRP, TN, TDN, TP and TDP were measured in a Flow Injection Analyser (Lachat Instruments, Colorado, USA). TOC and DOC were determined in the San++ Segmented Flow Analyser (Skalar, Breda, Netherlands). TS and TDS were analyzed in the 730ES Inductively Coupled Plasma–Atomic Emission Spectrometer (Agilent Technologies, California, USA). Principal component analysis PCA was performed using the PRIMER Version 6 statistical package (Clarke and RN, 2006) on the normalized physical and chemical parameters.

4.3.3 Epifluorescence microscopy

Water samples collected for microscopy were preserved in formaldehyde (1% v/v). Cells and VLP were vacuum filtered onto 25 mm polycarbonate 0.015 μm pore-size membrane filters (Nuclepore Track-etched, Whatman, GE Healthcare, USA) with a 0.45 μm pore-size backing filter. The 0.015 μm filter was mounted onto a glass slide with ProLong Gold anti fade reagent (Invitrogen, Life Technologies, NY, USA) and 2 μl (25 \times dilution in sterile filtered milliQ water <0.015 μm) SYBR Gold nucleic acid stain (Invitrogen, Life Technologies, NY, USA). Prepared slides were visualized in an epifluorescence microscope (Olympus BX61, Hamburg, Germany) under excitation with blue light (460–495

nm, emission 510–550 nm). Cell and virus-like particle (VLP) counts were performed on the same filter over 30 random fields of view.

4.3.4 Cellular diversity analyses

Diversity of *Bacteria*, *Archaea* and *Eucarya* was assessed using small subunit ribosomal RNA (SSU) gene sequences. Metagenomic reads that matched the 16S and 18S ribosomal RNA (rRNA) genes were retrieved using METAXA (Bengtsson *et al.*, 2011). Only sequences longer than 200 bp were accepted for downstream analysis.

The Quanitative Insights Into Microbial Ecology (QIIME) pipeline (version 1.4.0) (Caporaso *et al.*, 2010) implementing UCLUST, was used to group SSU sequences into operational taxonomic units (OTUs) at 97% percent identity against the SILVA SSU reference database (release 108) (www.arb-silva.de). SSU sequences that did not cluster with sequences from SILVA were allowed to form new OTUs (no suppression). A representative sequence from each OTU was chosen and classified to the genus level using QIIME implementing the Ribosomal Database Project (RDP) classifier (Wang *et al.*, 2007) trained against SILVA. Assignments were accepted to the lowest taxonomic rank with bootstrap value $\leq 85\%$.

To allow comparison of the relative abundance of taxa, the number of SSU matches per sample filter was normalized to the average number of reads (403,577). Statistical analysis on the relative SSU abundances was performed using the PRIMER version 6 package (Clarke and RN, 2006). The SSU counts of each sample filter were aggregated to the genus level and square root transformed to reduce the contribution of highly abundant taxa. A resemblance matrix was computed using Bray-Curtis similarity. The upper mixed zone (1.7, 4.2 and 5.7 m) and deep zone (6.5 and 6.7 m) samples were designated as separate groups and an analysis of similarity analysis of similarity (ANOSIM) performed to test for difference between the two groups. BEST analysis was performed with the abiotic variables: conductivity, temperature, turbidity, DO, pH, TOC, TN, TP, TS, total C:N, total C:P, total N:P, cell counts and VLP counts. The Bio-Env procedure in BEST looks at all the abiotic variables in combination and finds a subset sufficient to best explain the biotic structure. A heat map with bi-clustering dendrogram was generated using R and the package ‘seriation’ (Hahsler *et al.*, 2007) on the normalized square-root transformed SSU counts.

4.3.5 Analysis of functional potential

The relative abundance and taxonomic origin of functional marker genes was used to determine the potential for carbon, nitrogen and sulphur conversions. The ORFs were predicted from trimmed metagenomic reads using META-GENE (Noguchi *et al.*, 2006) accepting those >90 bp in length. Open reading frames ORFs were translated using the standard bacterial/plastid translation table and compared to protein sequences from the KEGG Genes database (release 58) using the basic local alignment search tool (BLAST) (Altschul *et al.*, 1990).

The BLAST output was processed using KEGG Orthology Based Annotation System (KOBAS) version 2.0 (Xie *et al.*, 2011) accepting assignments to KEGG Orthology (KO) groups with e-value <1e–05 and rank >5. KO groups used as functional markers are listed in Supplementary (Table 4.2). Marker enzymes were assigned to taxonomic groups based on the species of origin of the best KEGG Genes BLASTP match.

Table 4.2: Full list of KEGG orthologs involved in carbon, nitrogen and sulphur conversions that were searched for in the Organic lake metagenome.

Process	Gene	KO	Notes
C fixation	ribulose-bisphosphate carboxylase large (<i>cbbL</i>)	K01601	Calvin cycle
	ribulose-bisphosphate carboxylase small (<i>cbbS</i>)	K01602	Calvin cycle
	phosphoribulokinase (<i>prkB</i>)	K00855	Calvin cycle
	ATP-citrate lyase alpha (<i>aclA</i>)	K15230	rTCA cycle
	ATP-citrate lyase beta (<i>aclB</i>)	K15231	rTCA cycle
	citryl-CoA lyase (<i>ccl</i>)	K15234	rTCA cycle
	citryl-CoA synthetase (<i>ccsB</i>)	K15233	rTCA cycle
	carbon monoxide dehydrogenase/acetyl-CoA synthase alpha (<i>cdhA</i>)	K14138	WL
	carbon monoxide dehydrogenase/acetyl-CoA synthase beta (<i>cdhB</i>)	K00190	WL
Respiration	cytochrome C oxidase subunit I (<i>coxI</i>)	K02256	Eucaryotic
	cytochrome C oxidase subunit III (<i>coxIII</i>)	K02262	Eucaryotic
	cytochrome C oxidase subunit I (<i>coxA</i>)	K02274	Bacterial
	cytochrome C oxidase subunit III (<i>coxC</i>)	K02276	Bacterial
Fermentation	L-lactate dehydrogenase (<i>ldh</i>)	K00016	
	pyruvate:ferredoxin oxidoreductase alpha (<i>porA</i>)	K00169	
	pyruvate:ferredoxin oxidoreductase beta (<i>porB</i>)	K00170	
CO oxidation	carbon-monoxide dehydrogenase large (<i>coxL</i>)	K03520	
	carbon-monoxide dehydrogenase medium (<i>coxM</i>)	K03519	
	carbon-monoxide dehydrogenase small (<i>coxS</i>)	K03518	
AAnP	photosynthetic reaction center L (<i>pufL</i>)	K08928	
	photosynthetic reaction center M (<i>pufM</i>)	K08929	
Methanogenesis	coenzyme M methyl reductase (<i>mcrB</i>)	K00401	
	methyl coenzyme M reductase system	K00400	
CH ₄ oxidation	soluble methane monooxygenase	K08684	
	nitrogenase (<i>anfG</i>)	K00531	
N fixation	nitrogenase molybdenum-iron protein alpha (<i>nifD</i>)	K02586	
	nitrogenase iron protein (<i>nifH</i>)	K02588	
	nitrogenase molybdenum-iron protein beta (<i>nifK</i>)	K02591	
	nitric oxide reductase (<i>norB</i>)	K02305	
	nitric oxide reductase (<i>norC</i>)	K02305	
Denitrification	nitrous oxide reductase (<i>nosZ</i>)	K00376	
	periplasmic cytochrome c-552 (<i>nrfA</i>)	K03385	
	hydroxylamine oxidase (<i>hao</i>)	K10535	hzo-like
DNRA	glutamate dehydrogenase (<i>gudB, rocG</i>)	K00260	
	glutamate dehydrogenase (NAD(P)+)	K00261	
	glutamate dehydrogenase (<i>gdhA</i>)	K00262	
	assimilatory nitrate reductase	K00360	
	assimilatory nitrate reductase (<i>nasA</i>)	K00372	

Continued on next page

Table 4.2 – *Continued from previous page*

Process	Gene	KO	Notes
Nitrification	assimilatory nitrate reductase (<i>narG</i>)	K00367	
	glutamine synthetase (<i>glnA</i>)	K01915	
	glutamate synthetase (NADPH/NADH)(<i>gltB</i>)	K00265	
	glutamate synthetase (ferredoxin) (<i>gltS</i>)	K00284	
	ammonia monooxygenase subunit A (<i>amoA</i>)	K10944	
DSR	ammonia monooxygenase subunit B (<i>amoB</i>)	K10945	
	ammonia monooxygenase subunit C (<i>amoC</i>)	K10946	
	adenylylsulfatedoreductase subunit A (<i>aprA</i>)	K00394	SRB related
ASR	adenylylsulfatedoreductase subunit B (<i>aprB</i>)	K00395	SRB related
	sulfite reductase (<i>dsrA</i>)	K11180	SRB related
	sulfite reductase (<i>dsrB</i>)	K11181	SRB related
	adenylyl sulfate kinase (<i>cysC</i>)	K00860	
S mineralisation	sulfateadenylyltransferase (<i>cysN</i>)	K00956	
	sulfateadenylyltransferase (<i>cysD</i>)	K00957	
	cysteine diogenase (<i>cdoI</i>)	K00456	
DMSP reduction	thiosulfate/3-Mercaptopyruvate sulfurtransferase (<i>sseA</i>)	K01011	
	anaerobic dimethyl sulfoxidereductase A (<i>dmsA</i>)	K07306	

Marker genes not represented by a KO group were assessed by BLASTP queries of marker gene sequences with experimentally confirmed function (Table 4.3) against a database of translated ORFs predicted from metagenomic reads. Matches were accepted if the e-value was $<1e-10$ and sequence identity was within the range shared by homologs of the query sequence(s) (Table 4.3). Matches to marker genes were normalized to 100 Mbp per sample and counted. Normalized frequencies of markers from the same pathway were averaged and those from different pathways were summed.

The same marker genes and BLAST procedure was used to compare the DMSP catabolism and photoheterotrophy potential of Organic Lake with nearby Ace Lake (Lauro *et al.*, 2011), Southern Ocean (SO) (Wilkins *et al.*, 2012) and global ocean sampling (GOS) metagenomes (Rusch *et al.*, 2007). Counts of single copy gene *recA* were also determined to estimate the percentage of genomes containing each marker gene (percentage of marker genes relative to *recA*). Matches to *recA* were accepted with e-value $<1e-20$ according to the cut-off established by Howard *et al.* (2008). For GOS samples, the BLAST database was generated from peptide sequences retrieved from Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) (camera.calit2.net) while the other BLAST databases were produced as for Organic Lake. The total number of trimmed base pairs for GOS samples was estimated by multiplying the number of reads from each sample by the average read length (822 bp) (Rusch *et al.*, 2007).

Marker gene sequences for phylogenetic analysis were clustered using the CD-HIT web server (Huang *et al.*, 2010) at 90% global amino acid identity. A representative sequence from the clusters that resided within a desired conserved region and homologs from cultured strains were used in phylogenetic analyses performed in Molecular Evolutionary Genetic Analysis (MEGA) 5.05 (Tamura *et al.*, 2011) implementing MUS-

Table 4.3: Functional marker gene sequences used in this study as BLAST queries for retrieving homologues in the Organic Lake metagenomes. %ID, minimum amino acid identity for a match to be considered homologous.

Gene (%ID)	Organism	Accession	Reference
<i>dddD</i> (60)	<i>Marinomonas</i> sp.MWYL1	ABR72937.1	Todd <i>et al.</i> (2007)
	<i>Pseudomonas</i> sp.J465	ACY01992.1	Curson <i>et al.</i> (2010)
	<i>Psychrobacter</i> sp.J466	ACY02894.1	Curson <i>et al.</i> (2010)
	<i>Halomonas</i> sp. HTNK1	ACV84065.1	Todd <i>et al.</i> (2010)
<i>dddL</i> (45)	<i>Sulfitobacter</i> sp. EE36	ADK55772.1	Curson <i>et al.</i> (2008)
	<i>Rhodobacter sphaeroides</i> 2.4.1	ABA77574.1	Curson <i>et al.</i> (2008)
<i>dddP</i> (55)	<i>Roseovarius nubinhibens</i> ISM	EAP77700.1	Todd <i>et al.</i> (2009)
<i>dddQ</i>	<i>Ruegeria pomeroyi</i> DSS-3	AAV94883.1	
	<i>Roseovarius nubinhibens</i> ISM	EAP76001.1	
	marine metagenome	EAP76002.1	Todd <i>et al.</i> (2011)
		GOS_7860946	Todd <i>et al.</i> (2011)
		GOS_2632696	Todd <i>et al.</i> (2011)
		GOS_2469775	Todd <i>et al.</i> (2011)
<i>dddW</i>	<i>Ruegeria pomeroyi</i> DSS-3	AAV93771.1	Todd <i>et al.</i> (2012)
<i>dddY</i>	<i>Alcaligenes faecalis</i>	ADT64689.1	?
<i>dmdA</i> (50)	<i>Ruegeria pomeroyi</i> DSS-3	AAV95190.1	Howard <i>et al.</i> (2006)
	<i>Pelagibacter ubique</i> HTCC1062	YP_265671.1	Howard <i>et al.</i> (2006)
<i>rhodopsin</i>	<i>Dokdonia donghaensis</i> MED134	EAQ40507.1	Gómez-Consarnau <i>et al.</i> (2007)
	<i>Vibrio</i> sp. AND4	ZP_02194911.1	Gómez-Consarnau <i>et al.</i> (2010)
<i>pufL</i> (45)	<i>Salinibacter ruber</i> DSM 13855	YP_445623.1	Balashov <i>et al.</i> (2005)
	<i>Roseovarius tolerans</i>	ABK88229.1	Labrenz <i>et al.</i> (1999)
	<i>Congregibacter litoralis</i> KT71	ZP_01104363.1	?
<i>pufM</i> (45)	<i>Roseovarius tolerans</i>	ABK88230.1	Labrenz <i>et al.</i> (1999)
	<i>Congregibacter litoralis</i> KT71	ZP_01104362.1	?
<i>soxB</i> (45)	<i>Sulfurimonas denitrificans</i> DSM 1251	YP_392780.1	?
	<i>Thiomicrospira crunogena</i> XCL-2	ABB42141.1	?
<i>soxA</i> (45)	<i>Sulfurimonas denitrificans</i> DSM 1251	YP_392780.1	?
	<i>Thiomicrospira crunogena</i> XCL-2	YP_390871.1	?
<i>soxC</i> (45)	<i>Sulfurimonas denitrificans</i> DSM 1251	YP_394569.1	?
	<i>Thiomicrospira crunogena</i> XCL-2	YP_390427.1	?
<i>soxD</i> (45)	<i>Sulfurimonas denitrificans</i> DSM 1251	YP_394568.1	?
	<i>Thiomicrospira crunogena</i> XCL-2	YP_390427.1	?
<i>sqr</i> (60)	<i>Sulfurimonas denitrificans</i> DSM 1251	ABB43898.1	?
<i>recA</i>	<i>Escherichia coli</i> K12	P0A7G6.2	Howard <i>et al.</i> (2008)

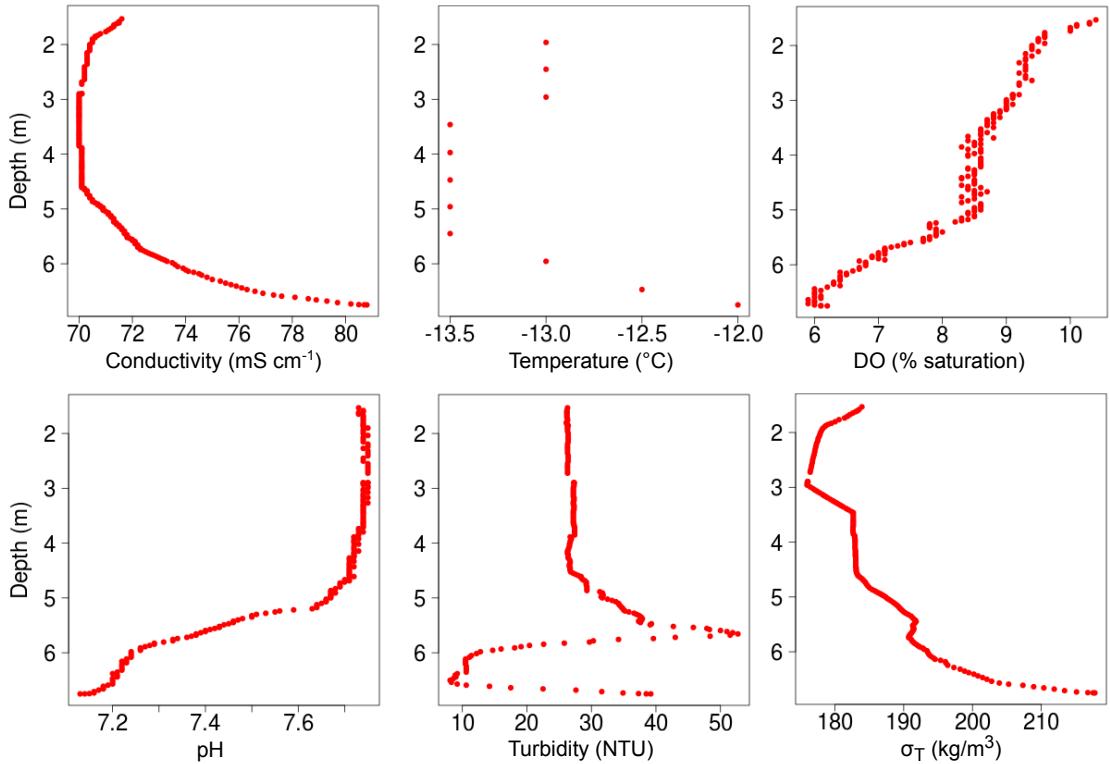


Figure 4.1: Vertical profiles of *in situ* Organic Lake abiotic parameters measured at the deepest point in the lake on 9 November 2008. $\sigma_T = (1000 - \text{density})$ was calculated from temperature and conductivity

CLE with default parameters (gap opening penalty: -2.9 , gap extension penalty: 0). Neighbor-joining was used to compute the phylogenies with a Poisson substitution model, uniform rates of change and complete deletion of alignment gaps. Node support was tested with bootstrap analysis (500 replicates).

4.4 Results and discussion

4.4.1 Abiotic properties and water column structure

In situ physico-chemical profiles (Figure 4.1) measured over the deepest point in the lake (Figure 4.2) determined the existence of two zones: an upper mixed zone above 5.7 m and a suboxic deep zone below 5.7 m (Figure 4.3). The separation of the two zones was indicated by a pycnocline and oxycline starting at 5.7 m. The pH also decreased with DO, likely due to fermentation products such as acetic, formic and lactic acids that have been reported in the bottom waters (Franzmann *et al.*, 1987b; Gibson *et al.*, 1994). The deep zone was not completely anoxic (Figure 4.1). Oxygen may be episodically introduced to bottom waters as a result of currents of cold dense water sinking during surface ice-formation (Ferris *et al.*, 1991). In comparison to meromictic lakes such as Ace Lake that have strong pycnoclines and a steep salt gradient in the anoxic zone, Organic Lake is shallow and has relatively weak stratification (Gibson, 1999).

Samples were collected from the upper mixed (1.7, 4.2 and 5.7 m) and deep (6.5 m

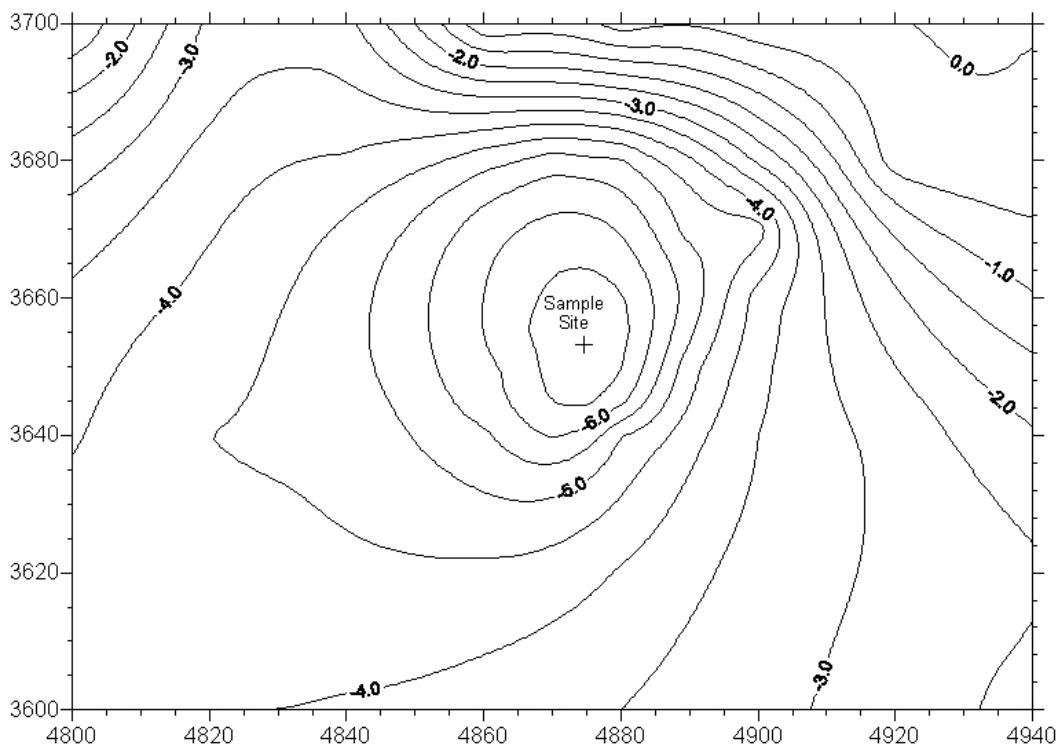


Figure 4.2: Bathymetry of Organic Lake 9 November 2008. Eastings and northings shown are abbreviated metric map co-ordinates.

and 6.7 m) zones. All nutrients, except for nitrate and nitrite reached maximum concentrations at 6.5 m (Table 4.4) suggestive of a layer of high biological activity above the lake bottom. Consistent with this, cell and VLP counts were highest at 6.5 m. However, turbidity was lowest at this depth demonstrating turbidity was not principally determined by cell density (Figure 4.3). Microscopy images did not show a shift in cell morphology that could account for the large drop in turbidity (Figure 4.4), which suggests particulate matter primarily contributed to turbidity readings. The low turbidity and peak in cell counts and nutrients at the oxycline at 6.5 m may be caused by an active microbial community degrading particulate matter. This inference is supported by the report of high concentrations of dissolved organic acids and free amino acids in the deep zone (Gibson *et al.*, 1994) as these nutrients are indicative of the breakdown of high molecular weight carbohydrates, lipids and proteins. Furthermore, the C:N and C:P ratios throughout the lake were high compared to the Redfield ratio (Redfield *et al.*, 1963) except at 6.5 m indicating this was the only depth where dissolved nitrogen and phosphorus were not relatively limited (Table 4.4).

Principal component analysis PCA of physico-chemical parameters showed all samples, except the 6.5 m sample, separated with depth along the PC1 axis (Figure 4.5). Accordingly, turbidity, TS and cell density were the strongest explanatory variables for the separation of the 6.5 m sample from the other deep sample, indicating that increased activity at 6.5 m was related to breakdown of particulate matter and sulphur chemistry.

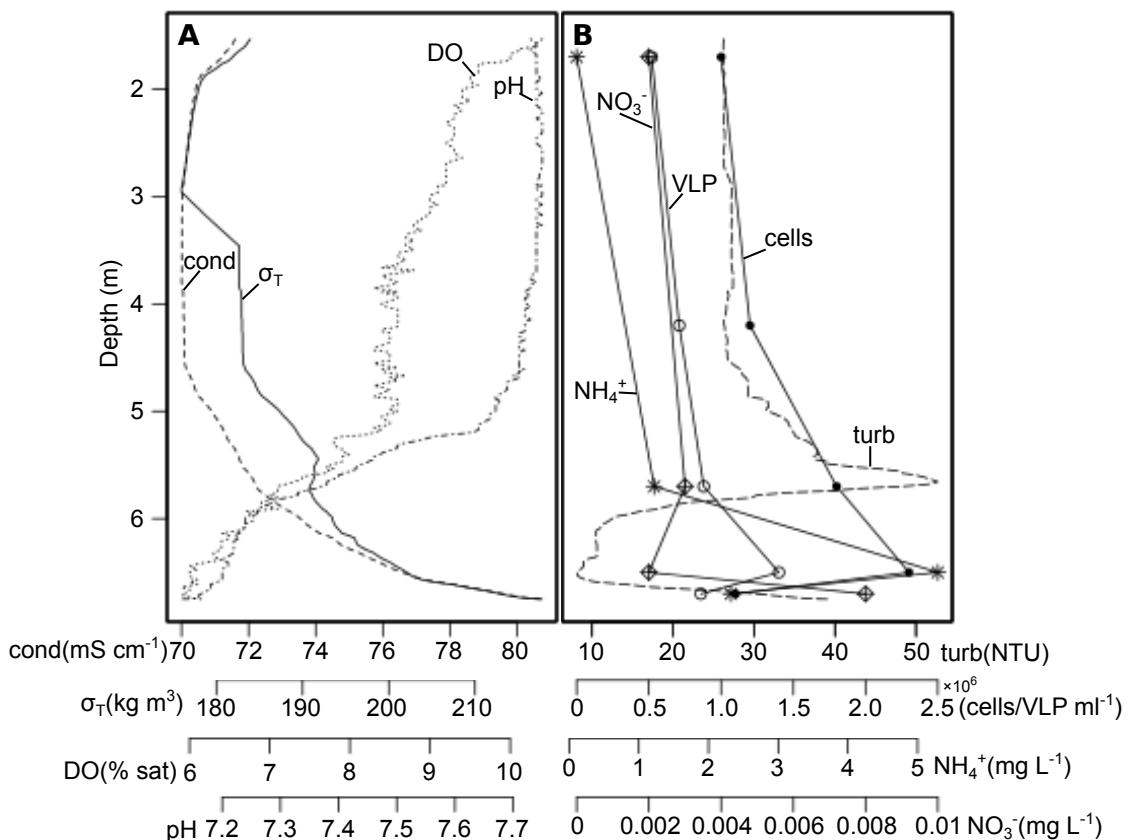


Figure 4.3: Vertical structure of Organic Lake. **(A)** Parameters that varied unimodally with depth showed two zones: an aerobic mixed zone above 5.7 m and a denser suboxic zone below. **(B)** Additional factors that revealed stratification within the deep zone. The peak in concentration at 6.5 m for ammonia was also observed for all other nutrients assayed except nitrate and nitrite, see (Table 4.4) for these values. $\sigma_T = (1000 - \text{density})$; cond, conductivity; DO, dissolved oxygen; turb, turbidity.

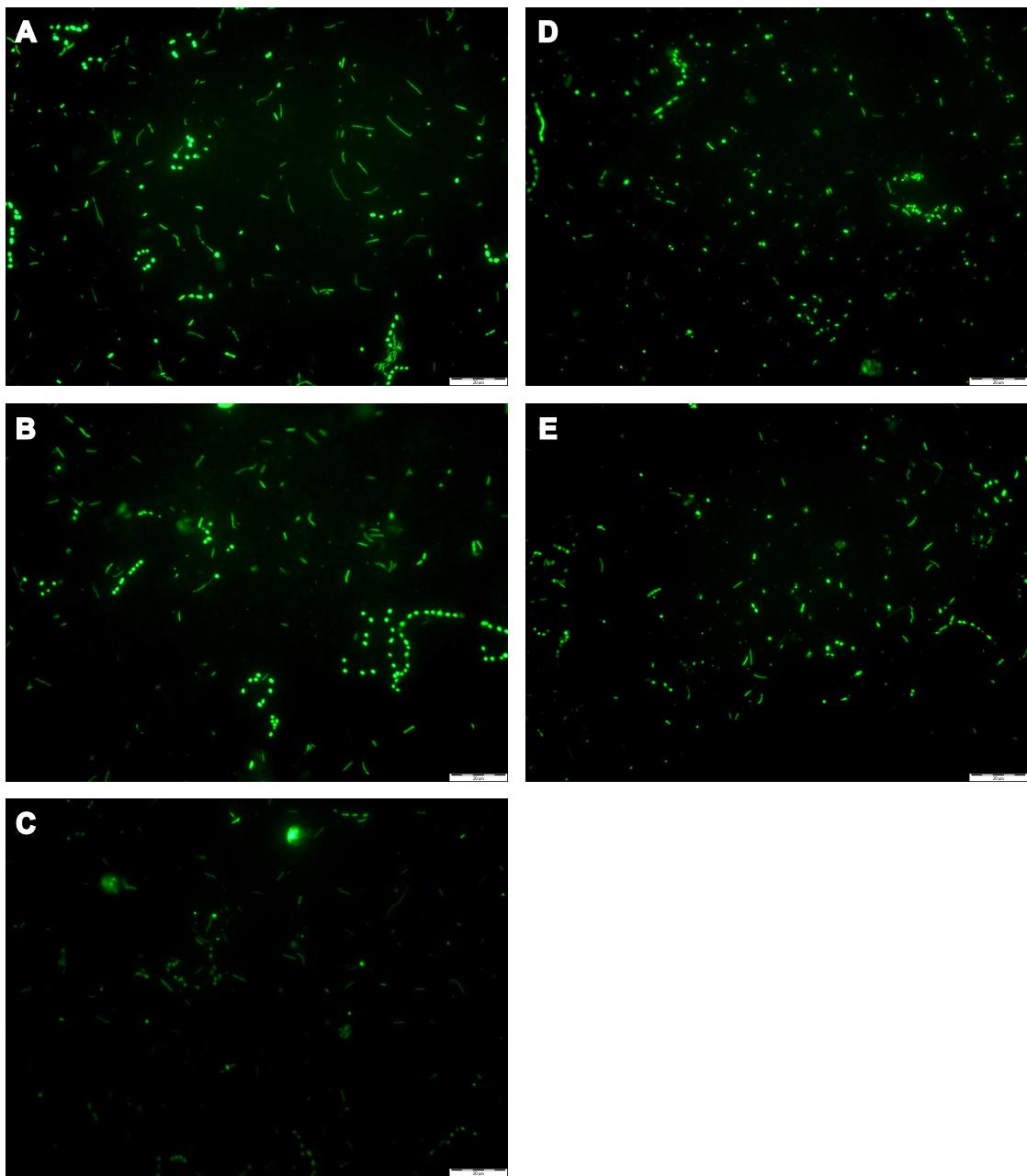


Figure 4.4: Epifluorescence microscopy images of Organic Lake microbiota ($<20\text{ }\mu\text{m}$) onto $0.015\text{ }\mu\text{m}$ polycarbonate membrane and stained with SYBR Gold. (A) 1.7 m, (B) 4.2 m, (C) 5.7 m, (D) 6.5 m, (E) 6.7 m. Scale bar = $20\text{ }\mu\text{m}$.

Table 4.4: Physico-chemical properties, cell counts and VLP counts of Organic Lake 2008 samples from a vertical profile. ND, data not determined.

	sample depths (m)				
	1.7	4.2	5.7	6.5	6.7
ammonia (mg l^{-1})	0.108	ND	1.22	5.29	2.32
nitrate (mg l^{-1})	<0.002	ND	0.003	<0.002	0.008
nitrite (mg l^{-1})	<0.002	ND	<0.002	0.010	0.010
DRP (mg l^{-1})	0.08	ND	0.10	0.20	0.18
TOC (mg l^{-1})	88	87	110	170	130
DOC (mg l^{-1})	69	ND	97	150	120
TN (mg l^{-1})	7.70	7.50	11	24	13
TDN (mg l^{-1})	0.112	ND	1.225	5.302	2.338
TP (mg l^{-1})	1.5	1.4	3.0	7.6	3.7
TDP (mg l^{-1})	0.509	ND	0.805	4.5	2
TS (mg l^{-1})	1010	974	1020	1410	950
TDS (mg l^{-1})	996	ND	1250	1290	995
particulate C:N:P (molar ratios)	49:7:1	ND	15:2:1	17:3:1	15:1:1
dissolved C:N:P (molar ratios)	350:20:1	ND	311:26:1	86:10:1	155:13:1
practical salinity	166	166	172	178	186
temperature ($^{\circ}\text{C}$)	-13	-13.5	-13	-12.5	-12
cells ml^{-1}	$1.0 \pm 0.4 \times 10^6$	$1.2 \pm 0.3 \times 10^6$	$1.8 \pm 0.5 \times 10^6$	$2.3 \pm 0.8 \times 10^6$	$1.1 \pm 0.4 \times 10^6$
VLP ml^{-1}	$5.2 \pm 2.1 \times 10^5$	$7.1 \pm 1.3 \times 10^5$	$8.8 \pm 3.4 \times 10^5$	$14 \pm 3.0 \times 10^5$	$8.6 \pm 3.3 \times 10^5$

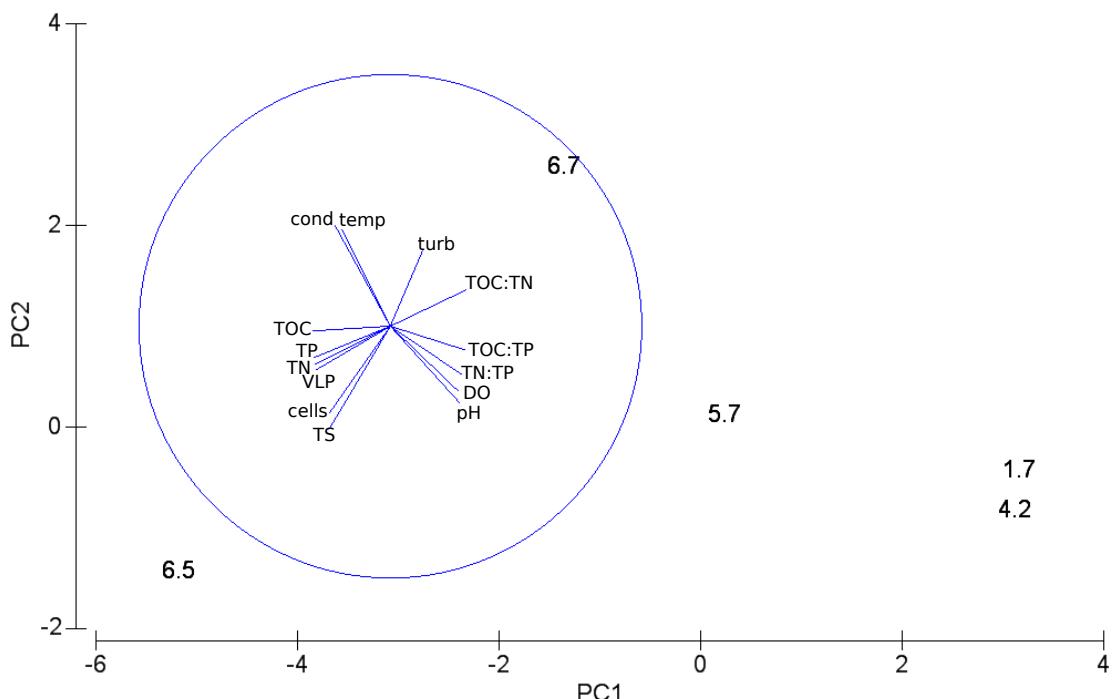


Figure 4.5: PCA analysis of physico-chemical parameters and cell/VLP counts of the Organic Lake profile. Data points are the sampling depths 1.7, 4.2, 5.7, 6.5 and 6.7 m. The overlaid vector diagram shows the relative contributions of the variables to explaining the difference between samples. PC1 explained 74.3% and PC2 explained 14.7% of the variation between samples. cond, conductivity; temp, temperature; turb, turbidity.

4.4.2 Overall microbial diversity

SSU genes (3,959 reads) that were retrieved from the metagenome data grouped into 983 OTUs. OTUs for *Bacteria* comprised 76.2%, *Eucarya* 16.3% and 7.5% of SSU sequences could not be classified. Only 2 reads, assigned to a deep sea hydrothermal clade of *Halobacteriales* (Supplementary Table S4), were assigned to *Archaea* indicating they were rare in Organic Lake.

The most abundant bacterial classes, *Gammaproteobacteria*, *Alphaproteobacteria* and *Flavobacteria*, were represented by OTUs on all filter sizes at all depths and each consisted of one dominant genus, *Marinobacter*, *Roseovarius* and *Psychroflexus*, respectively (Figure 4.6). Essentially all OTUs for *Cyanobacteria*/chloroplasts were classified as chloroplasts (Figure 4.6), except for three reads that could not be assigned to any lower rank (Supplementary Table S4) indicating free-living *Cyanobacteria* were rare or absent. OTUs for moderately abundant bacterial classes were *Actinobacteria*, *Deltaproteobacteria*, *Epsilonproteobacteria*, and candidate divisions OD1 and RF3. Lower abundance divisions included OTUs for *Bacilli*, *Clostridia*, *Spirochaetes*, *Lentisphaeria*, TM7, *Opitutae*, *Verrucomicrobia*, Bhi80-139, Bd1-5, SR1 and *Chlamydiae* (Figure 2A).

The dominant eucaryal OTUs were for photosynthetic *Chlorophyta* (green algae) and *Dictyochophyceae* (silicoflagellate algae) (Figure 4.7) principally assigned to the genus *Dunaliella* and the order *Pedinellales*, respectively (Supplementary Table S4). Lower abundance eucaryal OTUs included *Bacillariophyta* (diatoms), *Dinophyceae*, *Fungi* and heterotrophic *Choanoflagellida* and *Ciliophora* (see Supplementary Table S4 for lower taxonomic rank assignments).

4.4.3 Variation of microbial composition according to size and depth

Community composition varied with size fraction and depth. This was supported by seriation analysis that showed samples clustered according to size fraction, and those clusters further separated into upper mixed and deep zone groups (Figure 4.8). A significant difference in genus-level composition between the upper mixed and deep zones was supported by ANOSIM test (Rho: 0.53, significance: 0.1%). Differential vertical distribution of taxa is consistent with partitioning of ecological functions in the lake and in association with the physical and chemical data, described functional roles of those taxa.

20–3.0 µm fraction community composition

The upper mixed zone samples had a relatively high OTU abundance of *Dunaliella* chloroplasts and chlorophyte algae consistent with large active photosynthetic organisms concentrating near surface light. They are likely the main source of primary production in Organic Lake and have previously been reported to be the dominant algae (Franzmann *et al.*, 1987b). The SSU sequences for these algae at the bottom of the lake are likely to be due to sedimentation of dead cells or resting cysts.

Psychroflexus OTUs were overrepresented in the surface and 6.7 m samples. Con-

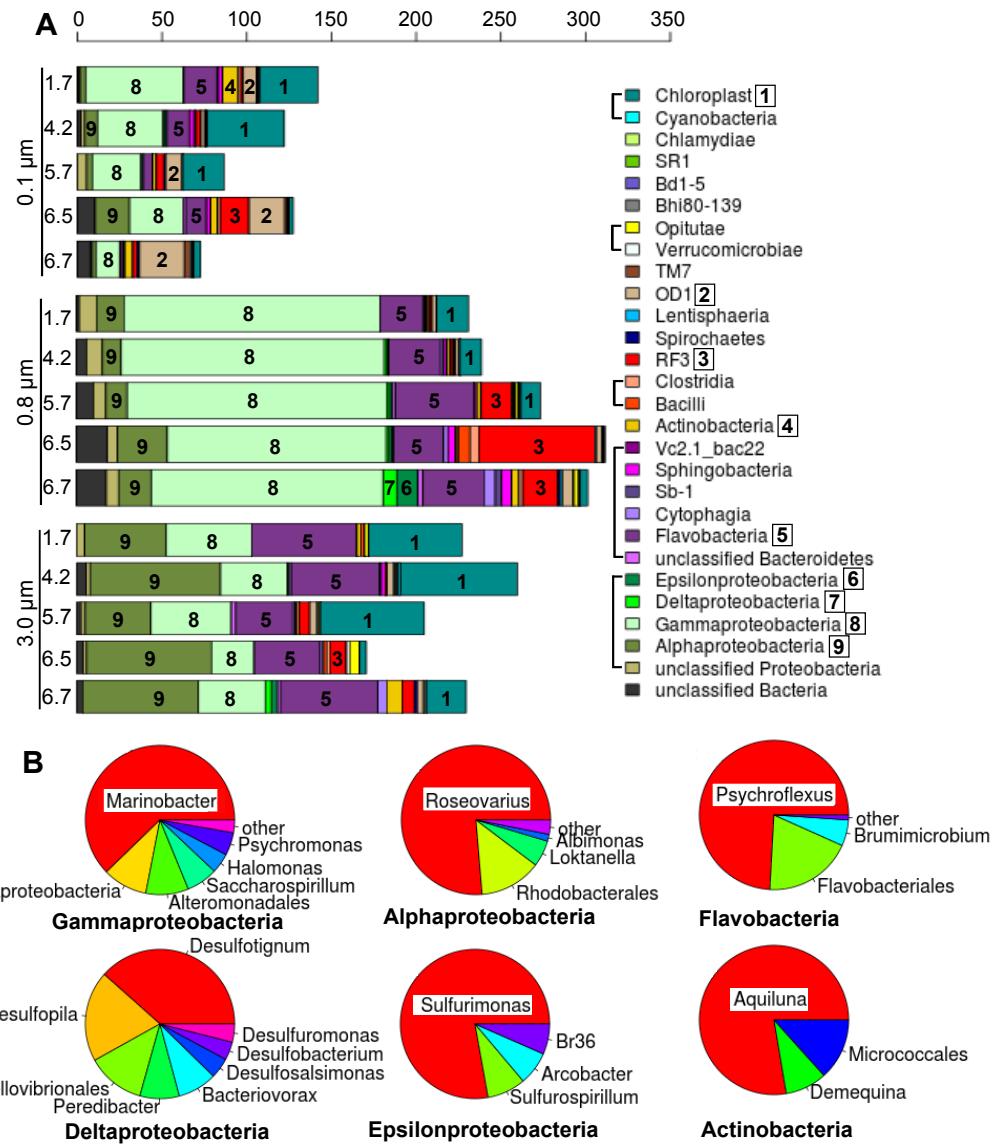


Figure 4.6: Diversity of (A) *Bacteria* from each size fraction (0.1, 0.8 and 3.0 μm) at each sample depth (1.7, 4.2, 5.7, 6.5 and 6.7 m) of Organic Lake aggregated according to class. The x-axis shows counts of SSU normalised to average reads acquired per sample filter. Taxa that belong to the same higher rank are shown grouped with a square bracket in the legend. Abundant taxa are labelled in plot with a number that corresponds to the numbered boxes in the legend. (B) Composition of abundant bacterial classes.

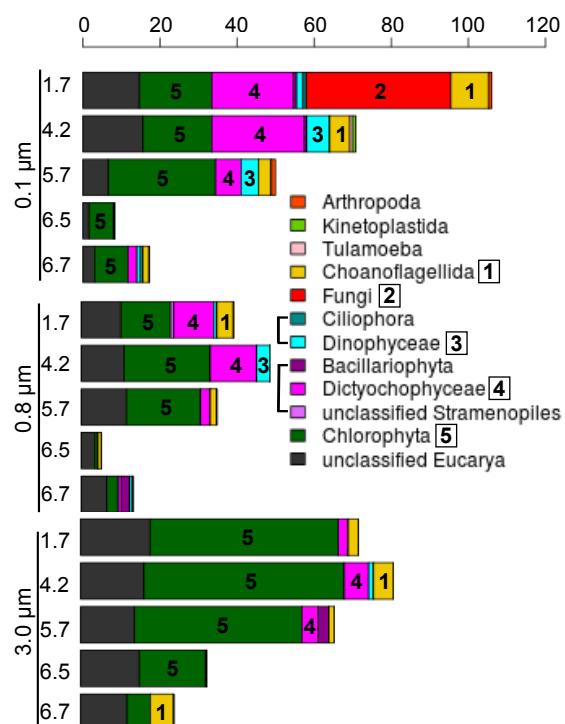


Figure 4.7: Diversity of *Eucarya* from each size fraction (0.1, 0.8 and 3.0 μm) at each sample depth (1.7, 4.2, 5.7, 6.5 and 6.7 m) of Organic Lake aggregated according to class. The x-axis shows counts of SSU normalised to average reads acquired per sample filter. Taxa that belong to the same higher rank are shown grouped with a square bracket in the legend. Abundant taxa are labelled in plot with a number that corresponds to the numbered boxes in the legend.

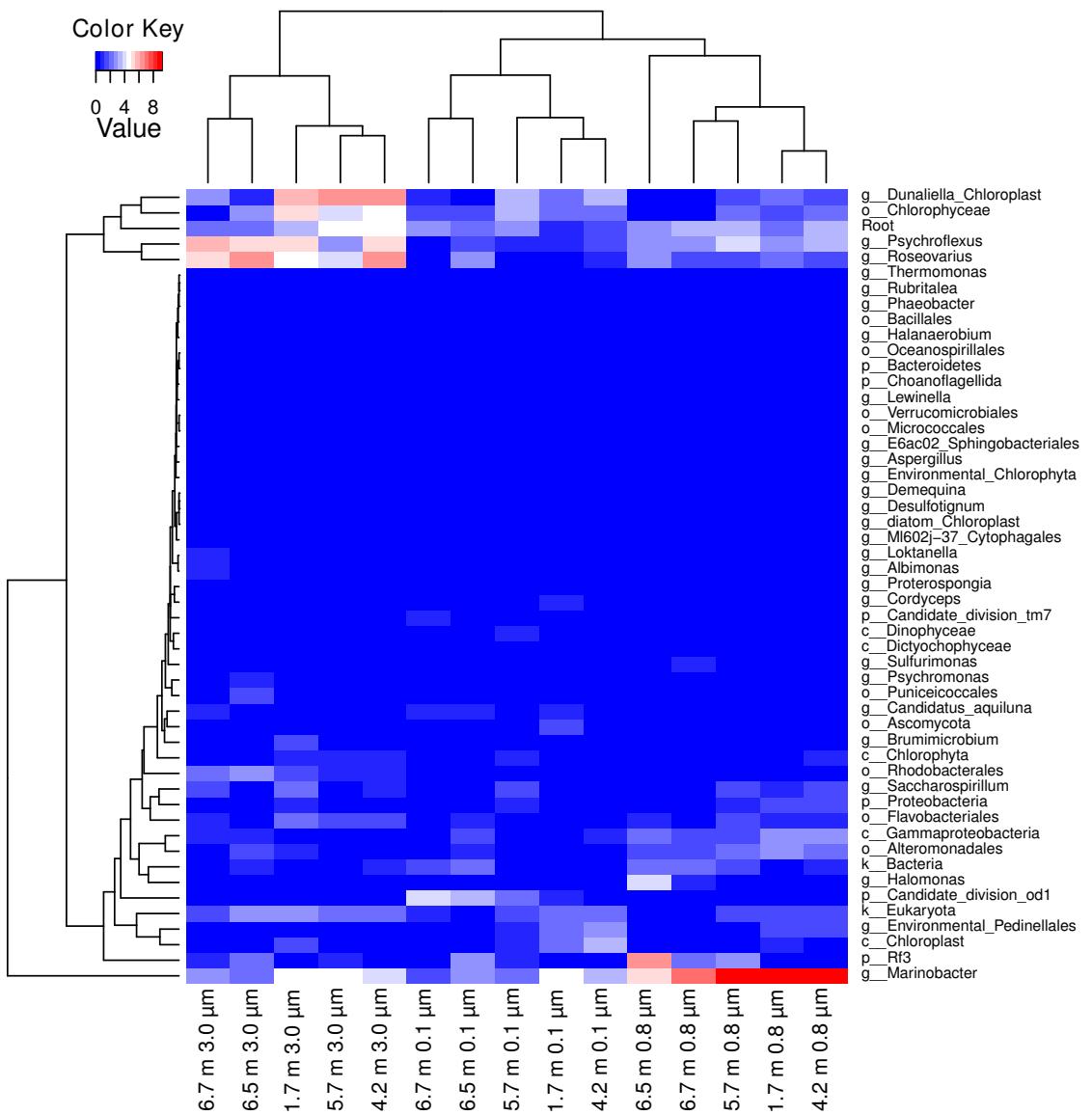


Figure 4.8: Heatmap and biplotting plot of the SSU gene composition in Organic Lake. Samples are shown according to size fraction (0.1, 0.8 and 3.0 μm) and depth (1.7, 4.2, 5.7, 6.5 and 6.7 m). SSU genes were classified to lowest taxonomic rank that gave bootstrap confidence $>85\%$ until the rank of genus. SSU gene counts were normalised and square root transformed. Taxa that comprised $<2\%$ of the sample were not included.

sistent with enrichment on the 3.0 μm filters, *Psychroflexus* (formerly *Flavobacterium*) *gondwanensis* (Bowman *et al.*, 1998) isolated from Organic Lake (Franzmann *et al.*, 1987b) had cells 1.5–11.5 μm in length (Dobson *et al.*, 1991). *Flavobacteria* associate with phytoplankton blooms in the Southern Ocean (Abell and Bowman, 2005a,b; Williams *et al.*, 2012), and have specialized abilities to degrade polymeric substances from algal exudates and detritus (reviewed in ?, (Williams *et al.*, 2012)). It is likely that Organic Lake *Psychroflexus* fills a similar ecological role. In support of this, *Psychroflexus* OTUs cluster with *Dunaliella* chloroplasts in the seriation analysis (Figure 4.8) and *P. gondwanensis* abundance in Organic Lake has been correlated with average hours of sunshine per day indicating population dynamics that is related to summer algal blooms (James *et al.*, 1994). The *Psychroflexus* OTUs in the deep zone are most likely due to sedimentation as *P. gondwanensis* non-motile and strictly aerobic (Dobson *et al.*, 1991).

Roseovarius OTUs were enriched at 4.2 m and 6.5 m suggesting different ecotypes may be present in the upper mixed zone compared to the deep zone. *Roseovarius tolerans*, an isolate from Ekho Lake in the Vestfold Hills, Antarctica has a cell size (1.1–2.2 μm ; (Labrenz *et al.*, 1999)) that would be expected to be captured on the 0.8 μm filter. The *Roseovarius* captured on the 3 μm filter may therefore be a different species, or a strain similar to *R. tolerans* from Ekho Lake that exhibits different growth characteristics (i.e. larger cell size or forms aggregates). A strain of this species from Ekho Lake is capable of microaerophilic growth (Labrenz *et al.*, 1999). Overrepresentation at 6.5 m may therefore be indicative of growth at that depth rather than sedimentation because sinking cells would be more abundant close to the lake bottom at 6.7 m. *Roseovarius* OTUs cluster with *Dunaliella* chloroplast and *Psychroflexus* OTUs in the seriation analysis (Figure 4.8), suggesting that Organic Lake *Roseovarius* may be utilising compounds released from algal-derived particulate matter, or made available by processing of complex organic matter by *Psychroflexus*. *Roseovarius* is a member of the *Roseobacter* clade, which is inferred to have an opportunistic ecology frequently associated with nutrient-replete plankton aggregates, including by-products of flavobacterial exoenzymatic attack (Moran *et al.*, 2007; ?). Additionally, the diverse metabolic capabilities of the *Roseobacter* clade include DMSP degradation, AAnP and CO oxidation (reviewed in Wagner-Döbler and Biebl (2006)). All of these capabilities should facilitate growth in both the upper mixed and deep zones of Organic Lake (see 4.4.5).

3–0.8 μm size fraction community composition

On the 0.8 μm filter, OTUs for *Marinobacter* dominated at all depths except 6.5 m. Their capture on this size fraction is consistent with the cell size of isolates (1.2–3 μm) (Gauthier *et al.*, 1992). The genus is metabolically versatile, which likely permits it to occupy the entire water column. *Marinobacter* is heterotrophic and the genus includes hydrocarbon-degrading strains (e.g., Gauthier *et al.* (1992); Huu *et al.* (1996), although deep-sea metal-oxidising autotrophs have also been reported (Edwards *et al.*, 2003). Some isolates are capable of interacting with diatoms (Gärdes *et al.*, 2010) and dinoflagellates (Green *et al.*, 2006). *Marinobacter* isolates from Antarctic lakes are capa-

ble of anaerobic respiration using dimethylsulphoxide (DMSO) (Matsuzaki *et al.*, 2006) or nitrate (Ward and Priscu, 1997). Analysis of functional potential linked to *Mari-nobacter* revealed additional metabolic capabilities potentially related to its dominance in Organic Lake (see Carbon resourcefulness in dominant heterotrophic bacteria and Molecular basis for unusual sulphur chemistry below).

OTUs for RF3 and Halomonas were overrepresented at 6.5 m, and RF3 sequences were more abundant (Figure 4.8). Their relative abundance in the deep zone indicates a role in microaerophilic processes. The majority of RF3 sequences to date are from anaerobic environments including mammalian gut (Tajima *et al.*, 1999; Ley *et al.*, 2006; Samsudin *et al.*, 2011), sediment (Yanagibayashi *et al.*, 1999; Röske *et al.*, 2012), municipal waste leachate (Huang *et al.*, 2005), anaerobic sludge (Chouari *et al.*, 2005; Goberna *et al.*, 2009; Rivière *et al.*, 2009; Tang *et al.*, 2011), a subsurface oil well head (Yamane *et al.*, 2011), and the anaerobic zone of saline lakes (Humayoun *et al.*, 2003; Schmidtova *et al.*, 2009; Bowman *et al.*, 2000a). However, some members have been found in surface waters (Demergasso *et al.*, 2008; Xing *et al.*, 2009; Yilmaz *et al.*, 2012) suggesting not all members are strict anaerobes.

Several *Halomonas* isolates have been sourced from Organic Lake including two described species *Halomonas subglaciescola* and *H. meridiana*, both of which grow as rods with dimensions consistent with capture on this size fraction (Franzmann *et al.*, 1987a; James *et al.*, 1990). Despite these isolates being aerobic, *Halomonas* has been reported to be enriched at the oxycline in Organic Lake (James *et al.*, 1994) indicating *Halomonas* in the lake plays an ecological role in the suboxic zone. This capacity may be linked to the ability of free amino acids and organic acids, which are abundant in the deep zone (Gibson *et al.*, 1994), to stimulate the growth of isolates (Franzmann *et al.*, 1987a).

0.8–0.1 µm size fraction community composition

A large number of eucaryal sequences were evident in the 0.1 µm size fraction. The upper zone was overrepresented by OTUs for *Pedinellales* (silicoflagellate algae) that co-varied with chloroplasts (Figure 4.8). *Pedinellales* have only been detected in Antarctic lakes from molecular studies (Unrein *et al.*, 2005; Lauro *et al.*, 2011) including Organic Lake (Yau *et al.*, 2011) (Chapter 3), and light microscopy studies of Antarctic Peninsula freshwater lakes reported 5–8 µm diameter cells resembling *Pseudopedinella* (Unrein *et al.*, 2005). It is possible that in Organic Lake small (0.80.1 µm) free-living members or chloroplast-containing cyst forms (Thomsen, 2007) exist. However, without evidence to support this (e.g. by microscopy) it seems more likely that the lake sustains a relatively small number of active photosynthetic cells and the sequences detected arise from cysts or degraded cellular material.

OTUs for *Candidatus Aquiluna*, in the Luna-1 cluster of *Actinobacteria* (Hahn *et al.*, 2004; Hahn, 2009) were most abundant at 1.7 m. The genus has small cells (<1.2 µm; (Hahn, 2009), accounting for their concentration on this size fraction. Although originally described in freshwater lakes, the same clade was detected in abundance in Ace Lake (Lauro *et al.*, 2011) and surface Artic seawater (Kang *et al.*, 2012) demonstrating

that they play ecological roles in polar saline systems. In Ace Lake surface waters they were associated with utilisation of labile carbon and nitrogen substrates (Lauro *et al.*, 2011), and in Organic Lake surface waters they probably perform similar functions. The presence of this clade in the deep zone implies a facultative anaerobic lifestyle or sedimented cells.

The bottom of the water column was distinguished by the presence of OTUs for candidate divisions OD1 and TM7. OD1 was more abundant, and its prevalence on this size fraction is consistent with similar findings for size fractionation of ground water (Miyoshi *et al.*, 2005). OD1 is consistently associated with reduced, sulphur-rich, anoxic environments (Harris *et al.*, 2004; Elshahed *et al.*, 2005). OD1 from Zodletone Spring, Oklahoma, was reported to possess enzymes related to those from anaerobic microorganisms (Elshahed *et al.*, 2005). Genomic analyses identified OTUs for OD1 in the anoxic zone of Ace Lake (Lauro *et al.*, 2011). The distribution of OD1 in Organic Lake is consistent with an anaerobic metabolism and potential involvement in sulphur chemistry.

4.4.4 Organic Lake functional potential

To determine the potential for functional processes in Organic Lake, gene markers for carbon, nitrogen and sulphur conversions were retrieved from metagenomic reads. BEST analysis showed that variation in the population structure was significantly correlated (Rho: 0.519, significance: 0.3%) with the abiotic parameters, DO, temperature, TS and TN. The DO gradient has an obvious effect of separating aerobic from anaerobic taxa, and allows oxygen sensitive nitrogen and sulphur processes to occur in the deep zone. Functional potential, taxonomic composition and the physico-chemical data were integrated to infer the carbon, nitrogen and sulphur cycles.

4.4.5 Carbon resourcefulness in dominant heterotrophic bacteria

In both the upper mixed and deep zones, potential for carbon fixation was much lower than for degradative processes, indicating potential for net carbon loss (Figure 4.9). Potential for carbon fixation via the oxygen-tolerant Calvin cycle (Figure 4.9) was originally assessed by presence of the marker genes ribulose-bisphosphate carboxylase oxygenase (RuBisCO) and phosphoribulokinase (*prkB*) (Hügler and Sievert, 2011). The majority of RuBisCO homologs were related to *Viridiplanteae* (Table 4.5) supporting the ecological role of green algae as the principle photosynthetic organisms.

RuBisCO was only associated with a small proportion of *Gammaproteobacteria* (Table 4.5), principally from sulphur-oxidising *Thiomicrospira*, indicating some *Gammaproteobacteria* are autotrophs. However, the majority of *prkB* matched to *Gammaproteobacteria* (Table 4.5), predominantly *Marinobacter*. Although deep-sea, iron-oxidising autotrophic members of *Marinobacter* have been isolated (Edwards *et al.*, 2003), all genomes reported for *Marinobacter* have *prkB* but lack RuBisCO genes. Across *Marinobacter* genomes the *prkB* homolog is consistently adjacent to a gene for a putative phosphodiesterase, suggesting that the enzymes expressed by these genes may be in-

Table 4.5: Contribution of different taxonomic groups to counts of marker genes involved in carbon conversions.

Taxon	Calvin cycle	prkB	Respiration	Fermentation	rTCA	WL	CO oxidation	AAnP
<i>Acidobacteria</i>	0	0	0.02	0	0	0	0	0
<i>Actinobacteria</i>	0	0	0.64	0.23	0	0	0.08	0
<i>Alphaproteobacteria</i>	0.05	0	4.84	0	0	0	6.74	6.98
<i>Aquificae</i>	0	0	0.06	0	0	0	0	0
<i>Bacteroidetes</i>	0	0	3.42	0	0	0	0	0
<i>Betaproteobacteria</i>	0.04	0.06	0.07	0.09	0	0	0.22	0
<i>Chlorobi</i>	0	0	0	0	0	0	0	0
<i>Chloroflexi</i>	0	0	0.02	0	0	0	0.07	0
<i>Chrysiogenetes</i>	0	0	0	0	0	0	0	0
<i>Cyanobacteria</i>	0.09	0	0	0	0	0	0	0
<i>Deferribacteres</i>	0	0	0.01	0	0	0	0	0
<i>Deinococcus-Thermus</i>	0.01	0	0.02	0	0	0	0	0
<i>Delta proteobacteria</i>	0	0	0.09	0	0	0.06	0.21	0
<i>Epsilonproteobacteria</i>	0	0	0	0	0.28	0	0	0
<i>Firmicutes</i>	0.01	0	0.01	4.90	0	0.02	0.15	0
<i>Fornicata</i>	0	0	0	0	0	0	0	0
<i>Fusobacteria</i>	0	0	0	0	0	0	0.03	0
<i>Gammaproteobacteria</i>	0.05	12.1	9.86	1.03	0	0	0.06	0.04
<i>Nitrospirae</i>	0	0	0	0	0	0	0	0
<i>Planctomycetes</i>	0	0	0.02	0.08	0	0	0	0
<i>Spirochaetes</i>	0	0	0	0.03	0	0	0.16	0
<i>Thermobaculum</i>	0	0	0	0	0	0	0	0
<i>Thermotogae</i>	0.01	0	0	0	0	0	0.17	0
<i>Verrucomicrobia</i>	0	0	0.13	0.05	0	0	0	0
<i>Crenarchaeota</i>	0	0	0	0	0	0	0.01	0
<i>Euryarchaeota</i>	0.04	0	0	0	0	0	0	0
<i>Alveolata</i>	0	0	0.03	0	0	0	0	0
<i>Euglenozoa</i>	0	0	0	0	0	0	0	0
<i>Opistokonta</i>	0	0	0.16	0	0	0	0	0
<i>Rhodophyta</i>	0.16	0	0.03	0	0	0	0	0
<i>Strameopiles</i>	0.34	0	0	0	0	0	0	0
<i>Viriplantae</i>	3.10	0.06	1.10	0	0	0	0	0

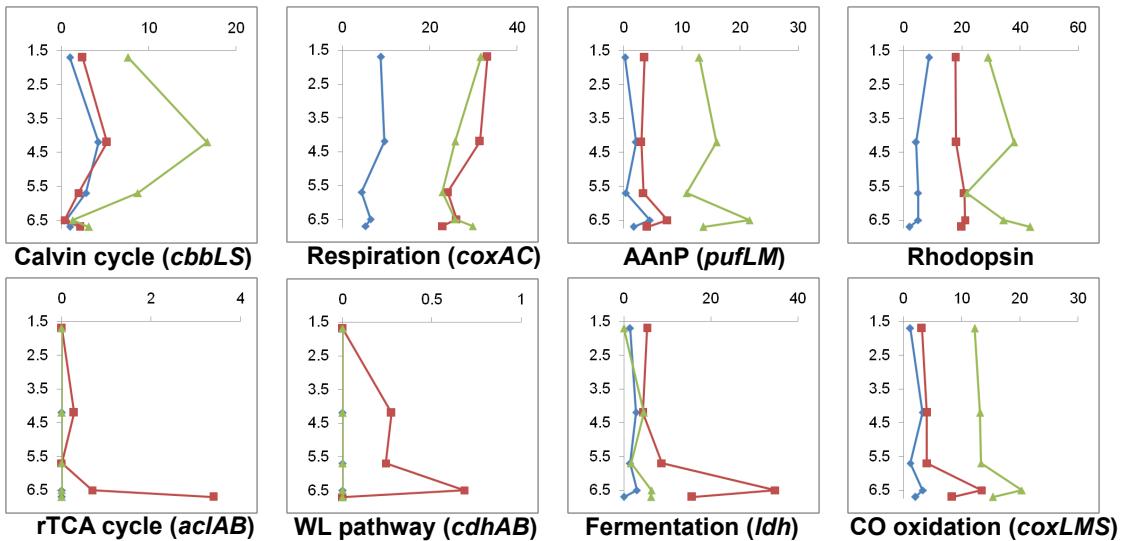


Figure 4.9: Vertical profiles of potential for carbon conversions for each size fraction in Organic Lake. The y-axis shows sample depths (m) and the x-axis shows counts of marker genes normalised to 100 Mbp of DNA sequence. The 0.1, 0.8, 3.0 μm size fractions are shown as blue, red and green, respectively. Counts for marker genes for the same pathway or enzyme complex were averaged and those from different pathways were summed. For marker gene descriptions see Table 4.2 and Table 4.3.

volved in a pathway involved in pentose phosphate metabolism unrelated to carbon fixation. Albeit exceptional, this decoupling of *prkB* from RuBisCO involved in carbon fixation (forms I and II), also observed in *Ammonifex* (Hügler and Sievert, 2011), undermines the utility of *prkB* as a marker gene for the Calvin cycle within certain groups. Thus, there is no evidence for autotrophy in Organic Lake mediated by *Marinobacter*.

Evidence for carbon fixation via the reverse tricarboxylic acid (rTCA) cycle was also indicated (Figure 4.9), with genes for ATP citrate lyase (*aclAB*) linked to sulphur-oxidising *Epsilonproteobacteria* (Table 4.5). In general, the rTCA cycle is restricted to anaerobic and microaerophilic bacteria (Hügler and Sievert, 2011), which is consistent with the detection of *Epsilonproteobacteria* in the lake bottom where oxygen is lowest, and the microaerophilic/anaerobic metabolisms characteristic of the group (Campbell *et al.*, 2006). Anaerobic carbon fixation was represented by potential for the Wood-Ljungdahl; or reductive acetyl-CoA (WL) pathway (Figure 4.9). WL-mediated carbon fixation, for which CO dehydrogenase/acetyl-CoA synthase is the key enzyme, was linked to *Firmicutes* and *Deltaproteobacteria* that are known to grow autotrophically using this pathway (Hügler and Sievert, 2011).

Potential for carbon loss by via respiration was indicated by an abundance of cytochrome C oxidase genes (*coxAC*) throughout the water column. In the deep zone, potential for fermentation was greatest at 6.5 m (Figure 4.9) and likely the main biological activity that was occurring at that depth. Fermentation was indicated by the marker gene lactate dehydrogenase (*ldh*). These genes were linked to *Firmicutes* (Table 4.5), which was only present at 6.5 m and represented by the classes *Clostridia* and *Bacilli* (Figure 4.6). As the related candidate division RF3 (Tajima *et al.*, 1999) also

has relatively high abundance in this zone (Figure 4.6) (see 0.8–3.0 μm size fraction community composition above), there is circumstantial evidence that RF3 possesses fermentative metabolism and may therefore play an important ecological role in Organic Lake by degrading high molecular weight compounds to organic acids that other organisms can utilize. Assimilation of fermentation products appears to play a greater role in Organic Lake rather than complete anaerobic oxidation involving methanogens or sulphate-reducing bacteria; the former were absent and the latter were present in low abundance (Figure 4.6).

Alphaproteobacteria, predominantly *Roseovarius* (Figure 4.6), were implicated in CO oxidation (Table 4.5), which is used to generate energy for lithoheterotrophic growth (Moran and Miller, 2007), although CO oxidation may also be involved in anaplerotic C fixation (Moran and Miller, 2007). The CO oxidation capacity was at a maximum at 6.5 m (Figure 4.9), and therefore associated with the deep-zone *Roseovarius* ecotype of Organic Lake. CO oxidation can function as a strategy to limit oxidation of organic carbon for energy so that a greater proportion can be directed towards biosynthesis (Moran and Miller, 2007).

Photosynthesis reaction center genes *pufLM*, involved in photoheterotrophy via AAnP, were abundant in Organic Lake (Figure 4.9, Table 4.5). These were linked to the *Roseobacter* clade of *Alphaproteobacteria* (Table 4.5), major contributors to AAnP in ocean surface waters (Béjà *et al.*, 2002; Moran and Miller, 2007). This is consistent with the known metabolic potential of bacteriochlorophyll A (BchlA) producing *Roseovarius tolerans* from Ekho Lake (Labrenz *et al.*, 1999). Photoheterotrophy can also be rhodopsin-dependent, with proteorhodopsin (PR) of marine *Flavobacteria* and *Vibrio* previously linked to light-dependent energy generation to supplement heterotrophic growth, particularly during carbon limitation (Gómez-Consarnau *et al.*, 2007, 2010). However, the function(s) of rhodopsins are diverse, and PRs are also hypothesized to be involved in light or depth sensing (Fuhrman *et al.*, 2008).

Rhodopsin genes were abundant in Organic Lake (Figure 4.9), and were associated with all the dominant Organic Lake aerobic heterotrophic lineages (Figure 4.10). Phylogenetic analysis revealed six well-supported Organic Lake rhodopsin groups (Supplementary Figure S6). All groups had an L or M residue at position 105 (vs the SAR86 PR), denoting tuning to surface green light (Man *et al.*, 2003; Gómez-Consarnau *et al.*, 2007), and is characteristic of oceanic coastal samples (Rusch *et al.*, 2007). Four of the groups clustered with homologs of genera detected in the lake, namely *Marinobacter*, *Psychroflexus*, *Octadecabacter* and “*Ca. Aquiluna*” (Figure 4.10) (Table S4). Another group (SAL-R group) originates from the sphingobacterium *Salinibacter ruber*, which produces xanthorhodopsin (Balashov *et al.*, 2005); it is therefore likely that Organic Lake *Sphingobacteria* (Supplementary Table S4) were the origin of this rhodopsin group. The most abundant group, OL-R1 (Figure 4.10) had no close homologs from GenBank, but it was abundant on the 3.0 μm fraction and has a distribution suggesting it originates from Organic Lake members of the *Roseobacter* clade (Figure 4.9). All ORFs adjacent to OL-R1 rhodopsin containing scaffolds were related to *Octadecabacter* further supporting

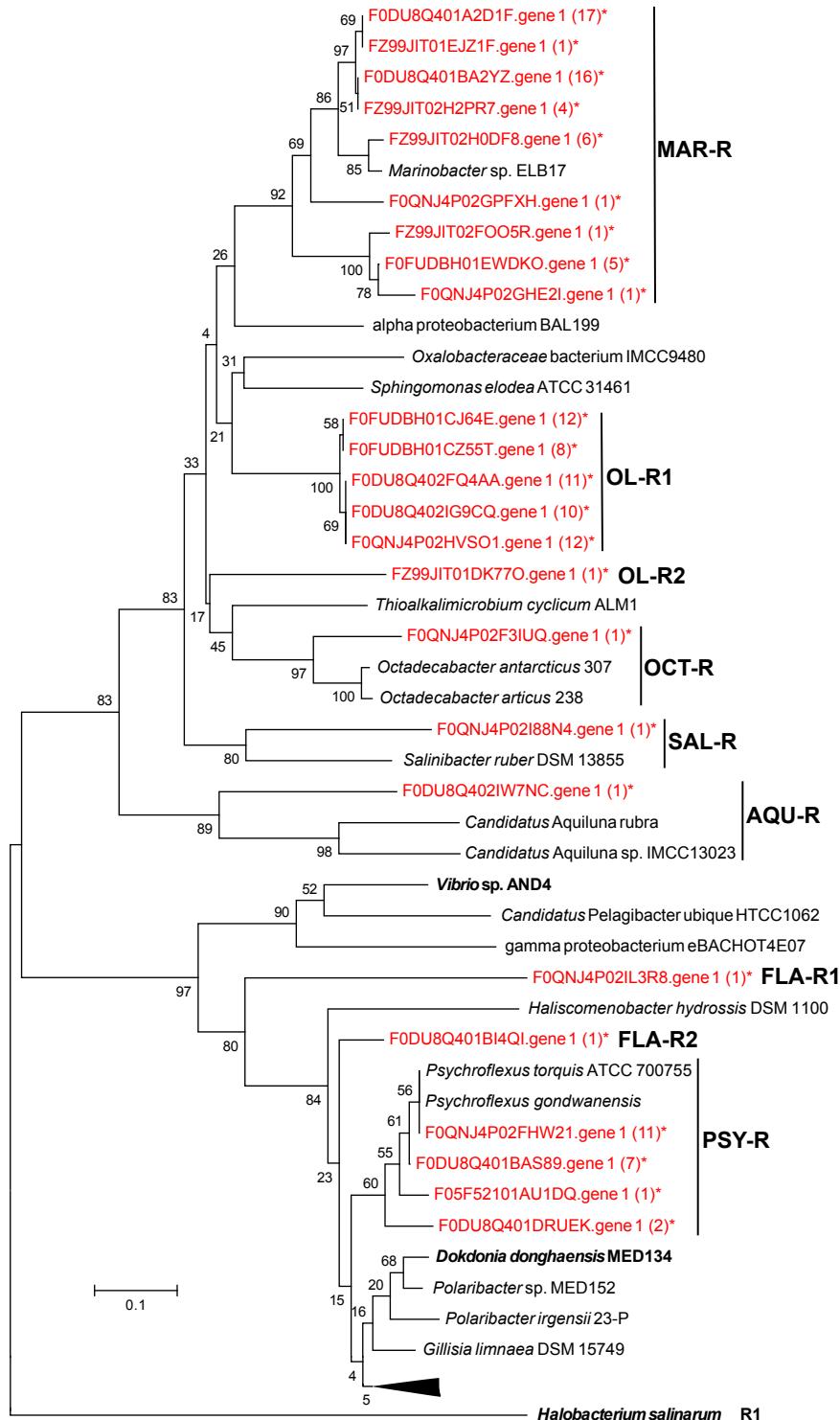


Figure 4.10: Phylogenetic tree of rhodopsin homologs. *Halobacterium salinarum* R1 halorhodopsin was used as an outgroup. The tree was computed from a 78 amino acid region spanning the motif involved in ‘spectral tuning’ using the neighbour-joining algorithm. Organic Lake sequences from this study are shown in red and marked with an asterisk (*). Numbers in parentheses are counts of sequences that clustered with the Organic Lake homologue shown in the tree with 90% amino acid identity. Sequences with confirmed activity are shown in bold. Accession numbers from top to bottom are: EAZ99241, EDP63929, EGF32634, ZP_09955974, AEG32267, EDY76405, EDY88259, YP_445623, ACN42850, EIC91904, ZP_02194911, AAZ21446, AAT38609, AEE49633, EAST1907, sequence from John Bowman (personal correspondence), EAQ40507, EAQ40925, EAR12394, EHQ04368 and YP_001689404.

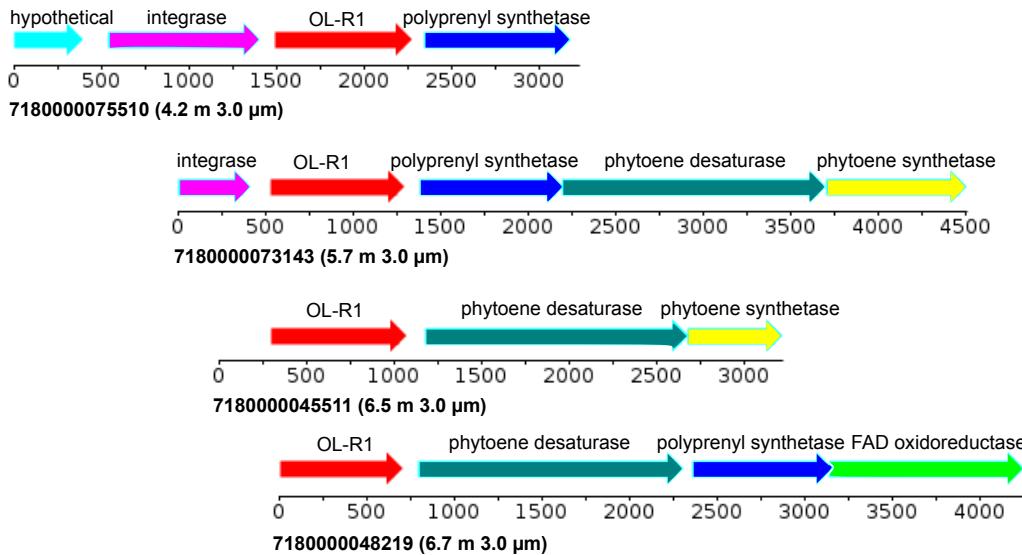


Figure 4.11: Genomic maps of Organic Lake scaffolds containing the OL-R1 rhodopsin homologue. All genes surrounding OL-R1 had best BLAST matches to *Octadecabacter* (*Alphaproteobacteria*) sequences. The scale below shows the number of base pairs. The sample depth and filter from which the scaffold was assembled is shown in parentheses beside the scaffold ID.

their *Roseobacter* clade provenance (Figure 4.11). Genes downstream of OL-R1 were involved in carotenoid synthesis, indicating OL-R1 is a xanthorhodopsin, occurring as a retinal protein or in a carotenoid complex (Balashov *et al.*, 2005).

Photoheterotrophic potential of Organic Lake was compared with other aquatic environments including nearby Ace Lake, SO and GOS expedition samples. The Organic Lake 0.1 μm fraction had the lowest rhodopsin counts and percentage of rhodopsin containing cells of all size-matched samples surveyed (Table 4.6). Non-marine GOS samples from the 0.1 μm fraction have been noted to have lower rhodopsin abundance (Sharma *et al.*, 2008), which was similarly evident from our analysis (Table 4.6). In contrast, the 3.0 μm Organic Lake size fractions had higher rhodopsin counts than Ace Lake and comparable counts to the SO samples, although the percentage of rhodopsin containing cells was still lower than that of the SO. The paucity of rhodopsins in the Organic Lake 0.1 μm fraction is likely due to the lack of SAR11 clade, which is expected to be the main source of rhodopsin genes in Ace Lake and marine samples. This indicates that although Organic Lake has an overall lower frequency of rhodopsin genes compared to sites for which size fraction-matched metagenomes are available, the rhodopsins associated with larger or particle-associated cells are as abundant as in the marine environment.

Table 4.6: Counts of genes involved in DMSP catabolism and photoheterotrophy in aquatic metagenomes (normalised to 100 Mbp). % = cells containing marker gene. The sample ID for each site is shown in parentheses after the site description. Values marked with an asterisk are >0 but <0.5 . Counts for the following sites are averages of several samples: Ace Lake mixolimnion (GS232, GS231); Southern Ocean SZ (GS349, GS351–GS353, GS356–GS360); Southern Ocean NZ (GS363, GS346, GS364, GS366GS368); GOS coastal (GS002–GS004, GS007–GS010, GS012–GS016, GS019, GS021, GS027–GS029, GS034–GS036); GOS open ocean (GS017, GS018, GS022, GS023, GS026, GS037, GS047); GOS estuary (GS006, GS011, GS012). Values shown in bold are the highest for that marker gene. SZ, Southern Zone; NZ, Northern Zone; GOS, Global Ocean Sampling.

Site	Size (μm)	<i>dddD</i> (%)	<i>dddL</i> (%)	<i>dddP</i> (%)	<i>dmdA</i> (%)	Rho. (%)	<i>pufLM</i> (%)	<i>recA</i>
Organic Lake 1.7 m (GS374)	0.1	2 (9)	4 (19)	0	0* (2)	1 (5)	0* (1)	21
	0.8	10 (36)	10 (39)	1 (2)	2 (7)	5 (20)	4 (14)	26
	3.0	11 (50)	5 (21)	2 (7)	9 (43)	12 (57)	13 (61)	21
Organic Lake 4.2 m (GS375)	0.1	5 (34)	5 (34)	0	1 (10)	1 (10)	2 (16)	14
	0.8	15 (54)	9 (31)	0	2 (6)	7 (23)	3 (11)	28
	3.0	23 (75)	2 (8)	1 (2.5)	20 (68)	14 (45)	16 (53)	30
Organic Lake 5.7 m (GS376)	0.1	4 (43)	1 (7)	0	1 (14)	2 (21)	0* (4)	10
	0.8	6 (20)	9 (32)	0	2 (7)	6 (22)	3 (12)	29
	3.0	19 (68)	3 (12)	0	13 (47)	6 (21)	11 (38)	28
Organic Lake 6.5 m (GS377)	0.1	10 (51)	0* (2)	0	3 (15)	1 (7)	4 (22)	20
	0.8	14 (38)	9 (23)	1 (2)	7 (20)	6 (16)	7 (20)	28
	3.0	42 (106)	5 (13)	0	20 (52)	6 (16)	22 (55)	29
Organic Lake 6.7 m (GS378)	0.1	1 (7)	0* (4)	0	0	1 (7)	2 (13)	13
	0.8	12 (26)	8 (17)	0	2 (5)	8 (16)	4 (9)	47
	3.0	50 (174)	5 (17)	4 (13)	12 (43)	12 (43)	14 (48)	29
Ace Lake mixolimnion	0.1	0* (2)	0	1 (2)	15 (56)	15 (53)	0* (1)	28
	0.8	2 (3)	1 (2)	0	2 (4)	12 (27)	3 (12)	45
	3.0	0	0	0* (4)	0	5 (42)	0	11
Newcomb Bay (GS235)	0.1	6 (14)	0	3 (7)	50 (111)	89 (196)	0	45
	0.8	5 (12)	0	0	18 (41)	55 (123)	0	45
	3.0	0	0	0	2 (17)	4 (33)	0	11
Southern Ocean SZ	0.1	2 (3)	0	6 (9)	71 (101)	98 (139)	0	70
	0.8	3 (6)	0* (0*)	5 (12)	32 (81)	43 (108)	0	39
	3.0	0* (7)	0	0* (4)	4 (66)	5 (84)	0	6
Southern Ocean NZ	0.1	0* (1)	0	5 (7)	124 (159)	111 (142)	1 (1)	78
	0.8	0* (2)	0	9 (30)	28 (84)	35 (107)	2 (7)	33
	3.0	0* (3)	0	1 (9)	7 (54)	11 (89)	0* (4)	12
GOS coastal	0.1	0* (0)	0	5 (6)	44 (52)	74 (87)	5 (6)	85
GOS open ocean	0.1	0	0	7 (8)	45 (50)	66 (74)	5 (5)	90
GOS estuary	0.1	0	0	1 (1)	29 (36)	61 (77)	2 (3)	80
GOS embayment (GS005)	0.1	4 (8)	0	6 (12)	28 (54)	58 (112)	3 (6)	52
GOS Lake Gatun (GS020)	0.1	0	0	0	4 (4)	48 (53)	2 (2)	90
GOS fringing reef (GS025)	0.1	0	0	0	0	7 (39)	0	18
GOS warm seep								

Continued on next page

Table 4.6 – *Continued from previous page*

Site	Size (μm)	<i>dddD</i> (%)	<i>dddL</i> (%)	<i>dddP</i> (%)	<i>dmdA</i> (%)	Rho. (%)	<i>pufLM</i> (%)	<i>recA</i>
(GS030) GOS upwelling	0.1	0	0	7 (6)	75 (63)	83 (69)	6 (5)	120
(GS031) GOS mangrove	0.1	0	0	4 (4)	81 (77)	81 (76)	4 (4)	106
(GS032) GOS Punta Cormorant	0.1	0	0	2(3)	24(34)	25(36)	1(1)	71
(GS033) GOS Rangirora Atoll (GS051)	0.1	0	11 (15)	14 (21)	4 (6)	31 (43)	15 (21)	72
	0.1	0	0	11(15)	38 (49)	73 (94)	3 (4)	77

Counts of *pufLM* genes in the Organic Lake 0.1 μm size fraction were similar to GOS sample, except for Punta Cormorant hypersaline lagoon which had the highest *pufLM* counts and percentage of AAnP cells (Table 4.6). However, the highest overall counts of *pufLM* were from the 3.0 μm size fraction of Organic Lake, likely due to the high proportion of members of the *Roseobacter* clade. Notably, *pufLM* genes were not detected in high abundance in Ace Lake or the SO samples, indicating AAnP is a unique adaptation in Organic Lake among these polar environments. The similarly high abundance of *pufLM* genes in Punta Cormorant hypersaline lagoon indicates AAnP may be advantageous in environments with salinity above marine levels.

The contribution of light-driven energy generation processes to the carbon budget is difficult to infer from genetic potential alone. For example, the relative abundance of AAnP and PR genes in Arctic bacteria has been reported to be the same in winter and summer (Cottrell and Kirchman, 2009). Furthermore, regulation of pigment synthesis is complex; for example, BchlA expression in *R. tolerans* occurs in the dark but is inhibited by continuous dim light (Labrenz *et al.*, 1999). However, it is possible that the apparent negative balance in carbon conversion potential could be ameliorated by photoheterotrophy performed by bacterial groups that are abundant in Organic Lake. In particular, the Organic Lake *Psychroflexus* could play a particular role as it has a PR related to *Dokdonia*, which was shown to function under carbon-limitation (Gómez-Consarnau *et al.*, 2007). Furthermore, detection of higher AAnP potential in Organic Lake than other aquatic environments linked with taxa known to be capable of AAnP, suggests it may have a greater influence in the carbon budget of Organic Lake.

4.4.6 Regenerated nitrogen is predominant in the nitrogen cycle

Nitrogen cycling potential throughout the lake was dominated by assimilation and mineralisation/assimilation pathways (Figure 4.12). Glutamate dehydrogenase (GDH) genes (*gdhA*) were abundant (Figure 4.12), and linked predominantly to *Alpha-* and *Gammaproteobacteria* and to a lesser extent *Bacteroidetes* (??). However, the functional significance of the readily reversible GDH depends on its origin; *Bacteroidetes* are likely to use GDH in the oxidative direction for glutamate catabolism (Williams *et al.*, 2012), whereas the use of GDH in the oxidative or reductive directions by *Pro-*

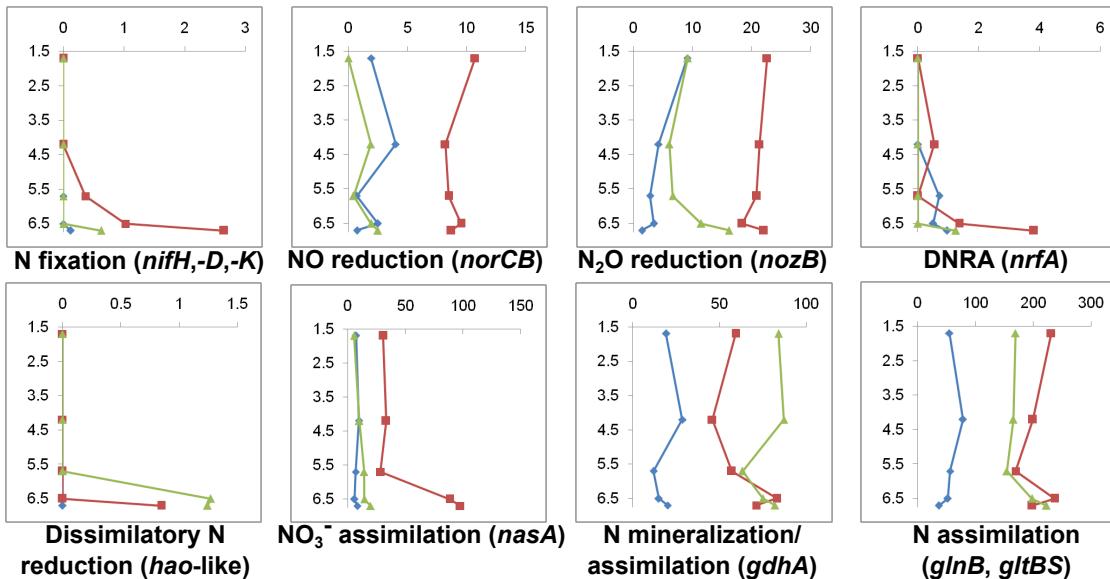


Figure 4.12: Vertical profiles of potential for nitrogen conversions for each size fraction in Organic Lake. The y-axis shows sample depths (m) and the x-axis shows counts of marker genes normalised to 100 Mbp of DNA sequence. The 0.1, 0.8, 3.0 μm size fractions are shown as blue, red and green, respectively. Counts for marker genes for the same pathway or enzyme complex were averaged and those from different pathways were summed. For marker gene descriptions see Table 4.2 and Table 4.3.

teobacteria is likely to depend upon the source of reduced nitrogen (ammonia vs amino acids). Glutamine synthetase (*glnB*) and glutamate synthase genes (*gltBS*), were predominantly linked to *Alpha-* and *Gammaproteobacteria* (??), indicating the potential for high-affinity ammonia assimilation by these groups in Organic Lake. The high ammonia concentration in the deep zone (Figure 4.3, Table 4.4) would result from a higher rate of mineralisation (ammonification) than assimilation. This is consistent with abundant OTUs for *Psychroflexus* (*Bacteroidetes*) in this zone, and due to either turnover of organic matter or lysis of *Bacteroidetes* cells after sedimentation in anoxic water. In addition, the gene for ammonia-generating nitrite reductase (*nrfA*) was linked to *Bacteroidetes* and *Planctomycetes* (??), indicating ammonia may also be produced by these putative aerobic heterotrophs. Overall, the data suggest that ammonia is actively assimilated in the aerobic upper mixed zone, but is permitted to accumulate in the anaerobic deep zone.

Potential for nitrogen conversions typically found in other aquatic environments was greatly reduced in Organic Lake. There was a very low potential for nitrogen fixation that was confined to the deep zone (Figure 4.12) and principally linked to anaerobic *Epsilonproteobacteria* (??). This diazotrophic potential may not be realized by *Epsilonproteobacteria*, given the high ammonia concentration present in the deep zone. No ammonia monooxygenase genes (*amoA*) were detected. The potential for ammonia oxidation was only represented by hydroxylamine/hydrazine oxidase-like (*hao*) genes, which were in low abundance and linked to *Deltaproteobacteria* (??). *hao* genes are present in non-ammonia-oxidising bacteria (Bergmann *et al.*, 2005), and those from Organic Lake belong to a family of multiheme cytochrome c genes present in sulphate-reducing

Table 4.7: Contribution of different taxonomic groups to counts of marker genes involved in nitrogen conversions.

Taxon	N fixation	NO reduction	N ₂ O reduction	DNRA	hao	N mineralisation	NO ₃ ⁻ assimilation	N assimilation
<i>Acidobacteria</i>	0	0	0	0	0.07	0	0.08	
<i>Actinobacteria</i>	0	0	0	0.03	0	0.32	0	5.41
<i>Alphaproteobacteria</i>	0.01	0.12	0	0	0	6.39	5.49	49.4
<i>Aquificae</i>	0	0	0.26	0	0	0	0	0.06
<i>Bacteroidetes</i>	0	0	3.00	0.27	0	3.90	0.03	15.5
<i>Betaproteobacteria</i>	0	0.03	0	0	0	0.06	0.41	19.2
<i>Chlorobi</i>	0.03	0	0	0	0	0.2	0	0.31
<i>Chloroflexi</i>	0	0	0	0	0	0.03	0	0.03
<i>Chrysiogenetes</i>	0	0	0	0	0	0.01	0	0.06
<i>Cyanobacteria</i>	0	0	0	0	0	0.29	0	0.10
<i>Deferribacteres</i>	0	0	0	0	0	0	0	0
<i>Deinococcus-Thermus</i>	0	0	0	0	0	0.01	0	0.11
<i>Deltaproteobacteria</i>	0.04	0.01	0	0.07	0.22	0.23	0	0.58
<i>Epsilonproteobacteria</i>	0.32	0	0	0	0	0.05	0	1.49
<i>Firmicutes</i>	0.03	0	0	0.03	0	0.70	0	3.16
<i>Fornicata</i>	0	0	0	0	0	0.02	0	0
<i>Fusobacteria</i>	0	0	0	0	0	0	0	0.04
<i>Gammaproteobacteria</i>	0	3.91	8.28	0	0	4.75	14.1	50.6
<i>Nitrospirae</i>	0	0	0	0.03	0	0.01	0	0
<i>Planctomycetes</i>	0	0.01	0	0.16	0	0	0	0.26
<i>Spirochaetes</i>	0	0.01	0	0	0	0.11	0	0.15
<i>Thermobaculum</i>	0	0	0	0	0	0.10	0	0
<i>Thermotogae</i>	0	0	0	0	0	0	0	0
<i>Verrucomicrobia</i>	0	0	0.17	0.03	0	0.25	0	0.82
<i>Crenarchaeota</i>	0	0	0	0	0	0.02	0	0
<i>Euryarchaeota</i>	0	0	0	0	0	0.12	0.09	0.10
<i>Alveolata</i>	0	0	0	0	0	0.18	0	0.03
<i>Euglenozoa</i>	0	0	0	0	0	0	0	0
<i>Opistokonta</i>	0	0	0	0	0	0.15	0	0.13
<i>Rhodophyta</i>	0	0	0	0	0	0	0	0.02
<i>Strameopiles</i>	0	0	0	0	0	0.03	0	0.15
<i>Viridiplantae</i>	0	0	0	0	0	0.03	0	0.35

*Delta*proteobacteria that have no proven role in ammonia oxidation. In the genomes of sulphate-reducing *Delta*proteobacteria thehao gene is invariably situated adjacent to a gene for a NapC/NirT protein, which suggests a role in dissimilatory nitrate reduction. Collectively these data indicate an inability for nitrification to occur in the upper mixed zone and no potential for ammonia loss in the deep zone.

Denitrification genes (*norCB* and *nozB*) and genes for nitrate assimilation (*nasA*) were present throughout the water column (Figure 4.12) and were linked primarily to *Gammaproteobacteria* (??). Low nitrate and nitrite in the deep zone (Figure 4.3, Table 4.4) indicates oxidized nitrogen has been depleted by dissimilatory or assimilatory reduction by heterotrophic *Gammaproteobacteria*. Denitrification genes are phylogenetically widespread and usually induced by low oxygen or oxidized nitrogen species (Kraft *et al.*, 2011) and thus expected to be active in the deep zone or oxycline. However, denitrification may be inhibited even if conditions appear appropriate. For example, in Lake Bonney, Antarctica, denitrification occurs in the west lobe, but not in the east lobe of the lake despite the presence of anoxia, nitrate and denitrifying *Marinobacter* species (Ward and Priscu, 1997; Ward *et al.*, 2005). Moreover, in the absence of nitrification, denitrification and nitrate assimilation would be limited by the lack of potential to re-form oxidized nitrogen. The preponderance of assimilation/mineralisation pathways geared towards reduced nitrogen appears to reflect a “short circuit” of the typical nitrogen cycle that would conserve nitrogen in a largely closed system. Hence, the predominant nitrogen source is regenerated fixed nitrogen. Similar findings were also made for Ace Lake, although in this system the presence of a dense layer of green sulphur bacteria with the potential to fix nitrogen augments the nitrogen cycle (Lauro *et al.*, 2011).

4.4.7 Molecular basis for unusual sulphur chemistry

Several meromictic hypersaline lakes in the Vestfold Hills, including Organic Lake, with practical salinity >150 are characterized by an absence of hydrogen sulphide and photoautotrophic sulphur bacteria (Burke and Burton, 1988). Although sulphate is present (Franzmann *et al.*, 1987b), geochemical conditions of these lakes are not conducive to dissimilatory sulphur cycling between sulphur oxidising and sulphate reducing bacteria typical of other stratified systems such as Ace Lake (Ng *et al.*, 2010; Lauro *et al.*, 2011). Consistent with this, potential for dissimilatory sulphate reduction represented by dissimilatory sulfite reductase (*dsrAB*) and adenylylsulphate reductase (*aprAB*) linked to sulphate-reducing *Delta*proteobacteria (Table 4.8) was low in Organic Lake. Sulphate-reduction potential was confined to the 6.7 m sample (Figure 4.13) where oxygen concentration was lowest and *Delta*proteobacteria were present (Figure 4.6).

Capacity for oxidation of reduced sulphur compounds, represented by the sulphur oxidation multienzyme genes (*soxAB*), was present throughout the water column (Figure 4.13) and linked primarily to *Alpha*- and *Gammaproteobacteria* (Table 4.8). Sulphur-oxidising *Alpha*- and *Gammaproteobacteria* are known to oxidise sulphur compounds, such as thiosulphate, aerobically. Although a small proportion of *Gammaproteobacteria* had the capacity for autotrophy (see 4.4.5), the majority of sulphur-oxidizers

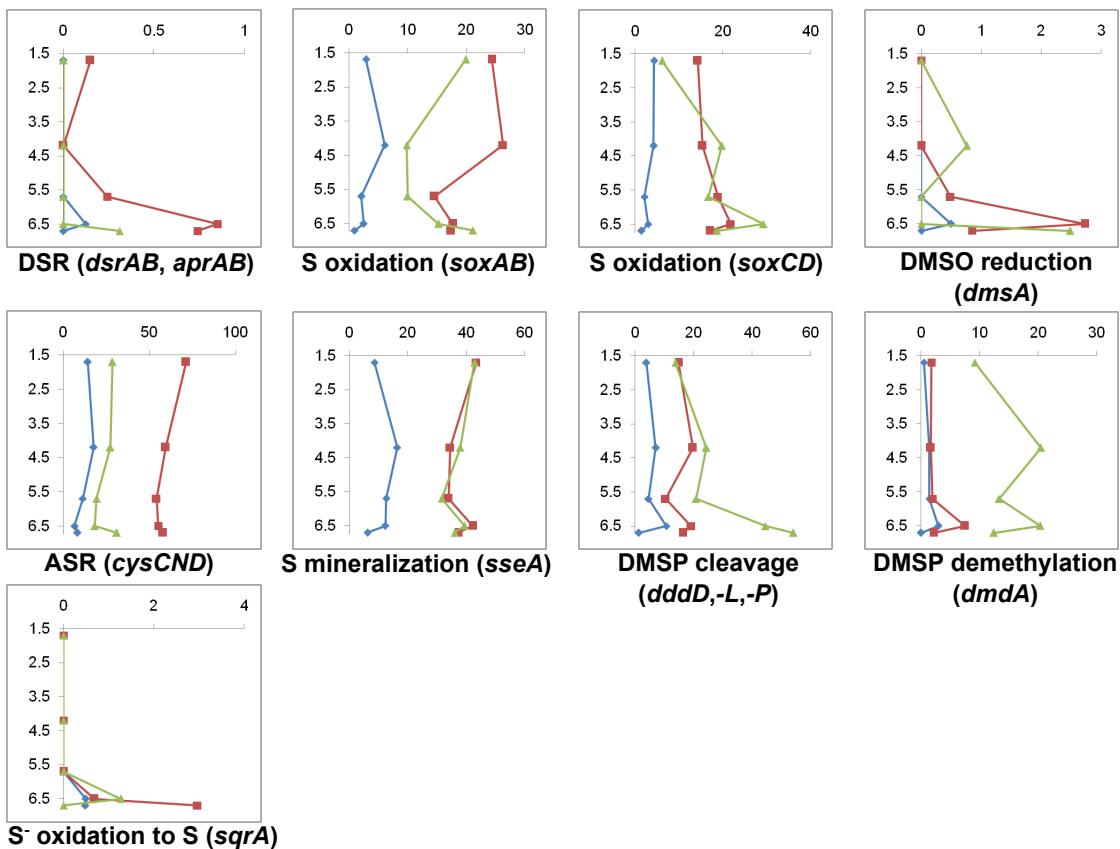


Figure 4.13: Vertical profiles of potential for sulphur conversions for each size fraction in Organic Lake. The y-axis shows sample depths (m) and the x-axis shows counts of marker genes normalised to 100 Mbp of DNA sequence. The 0.1, 0.8, 3.0 μm size fractions are shown as blue, red and green, respectively. Counts for marker genes for the same pathway or enzyme complex were averaged and those from different pathways were summed. For marker gene descriptions see Table 4.2 and Table 4.3.

were likely chemolithoheterotrophs as they were related to heterotrophic *Marinobacter* and *Roseobacter* clade. The sulphur dehydrogenase genes *soxCD* linked to *Alpha-* and *Gammaproteobacteria* were similarly present throughout the water column. *soxCD* are accessory components of the Sox enzyme system without which complete oxidation of thiosulphate cannot occur (Friedrich *et al.*, 2005). Thus the presence of *soxCD* indicates complete oxidation likely occurs, although the different distribution of *soxAB* and *soxCD* in the water column (Figure 4.13) suggests a proportion of the community may lack *soxCD* and deposit sulphur.

Sulphur-oxidising *Epsilonproteobacteria* possessing *soxAB* genes (Table 4.8) were present only in the deep zone of Organic Lake (Figure 4.6) and were related to autotrophic deep sea sulphur-oxidisers, some members of which are capable of anaerobic sulphur oxidation using nitrate (Yamamoto and Takai, 2011). It is unlikely that appreciable sulphur oxidation occurs in the deep zone as the known terminal electron acceptors, oxygen and nitrate, are deplete and the abundance of sulphur oxidising *Epsilonproteobacteria* is low (Figure 4.6). *Epsilonproteobacteria* were also linked to a capacity for oxidation of sulphide to elemental sulphur by utilising sulphide:quinone oxidoreductase (*sqrA*) (Figure 4.13, Table 4.8). In this pathway, sulphur is released as polysulphides, which is a potential biological source of the abundant polysulphides that have been detected in the lake (Roberts *et al.*, 1993).

It is likely that the limited anaerobic dissimilatory sulphur cycle contributes to the accumulation of DMS in Organic Lake in the deep zone. In the upper mixed zone, DMS could potentially be oxidized as a carbon and energy source or utilized as an electron donor by sulphur-oxidising bacteria (Schäfer *et al.*, 2010). In anoxic zones, methanogenic *Archaea* or sulphate-reducing bacteria are the main organisms known to breakdown DMS (Schäfer *et al.*, 2010). Methanogens and genes involved in methanogenesis were not detected, nor has methane been detected (Gibson *et al.*, 1994) leaving sulphate-reduction the most likely route of DMS catabolism. The low dissimilatory sulphate reduction potential in the deep zone coupled with the relatively stagnant waters would likely minimize DMS oxidation and loss by ventilation. DMS would therefore be expected to accumulate in the deep zone if production rates were higher than breakdown.

To determine the source of high DMS in the bottom waters of Organic Lake, the genes involved in DMS formation were surveyed. Genes for DMSP lyases *dddD*, *dddL* and *dddP*, were detected in Organic Lake at levels comparable to other dominant processes such as respiration and fermentation (Figure 4.13) indicating DMSP is an important carbon and energy source in Organic Lake. *dddD* was the most abundant of the Organic Lake DMSP lyases (Table 4.6) and comprised two main types: MAR-*dddD* and OL-*dddD* (Figure 4.14). Neither of these types clustered with the non-functional *Dinoroseobacter shibae* DFL 12 and *Ruegeria pomeroyi* DSS-3 *dddD* homologs (Todd *et al.*, 2011) or carnitine coenzyme A transferase outgroups, thereby providing support for their proposed role as functional DMSP lyases. The MAR-*dddD* type includes the *Marinobacter* sp. ELB17 *dddD* homolog, and MAR-*dddD* sequences were most abundant on the 0.8 µm fraction where *Marinobacter* OTUs were also abundant, indicating MAR-*dddD* derives

Table 4.8: Contribution of different taxonomic groups to counts of marker genes involved in sulphur conversions.

Taxon	DSR	S oxidation sqrA	S assimilation	S mineralisation	DMSO reduction
<i>Acidobacteria</i>	0	0	0.03	1.03	0
<i>Actinobacteria</i>	0	0	0.15	0	0
<i>Alphaproteobacteria</i>	0	2.05	0	0.85	11.7
<i>Aquificae</i>	0	0	0	0	0
<i>Bacteroidetes</i>	0	0	5.06	0.20	0
<i>Betaproteobacteria</i>	0	1.09	0	2.07	0.69
<i>Chlorobi</i>	0	0	0	0.03	0.15
<i>Chloroflexi</i>	0	0	0	0	0.30
<i>ChrysioGenetes</i>	0	0	0	0.01	0
<i>Cyanobacteria</i>	0	0	0	0.13	0.05
<i>Deferribacteres</i>	0	0	0	0	0
<i>Deinococcus-Thermus</i>	0	0	0	0	0.09
<i>Delta proteobacteria</i>	0.19	0	0	0.56	0.20
<i>Epsilonproteobacteria</i>	0	0.03	0.39	0.13	0
<i>Firmicutes</i>	0	0	0	0.22	0.09
<i>Fornicata</i>	0	0	0	0	0
<i>Fusobacteria</i>	0	0	0	0.03	0.13
<i>Gammaproteobacteria</i>	0	2.64	0	22.4	14.0
<i>Nitrospirae</i>	0	0	0	0	0
<i>Planctomycetes</i>	0	0	0	0.03	0
<i>Spirochaetes</i>	0	0	0	0	0.03
<i>Thermobaculum</i>	0	0	0	0	0.08
<i>Thermotogae</i>	0	0	0	0.01	0
<i>Verrucomicrobia</i>	0	0	0	0.18	0
<i>Crenarchaeota</i>	0.02	0	0	0	0
<i>Euryarchaeota</i>	0	0	0	0.02	0.12
<i>Alveolata</i>	0	0	0	0	0
<i>Euglenozoa</i>	0	0	0	0	0.03
<i>Opistokonta</i>	0	0	0	0.11	0.03
<i>Rhodophyta</i>	0	0	0	0	0
<i>Strameopiles</i>	0	0	0	0	0
<i>Viridiplantae</i>	0	0	0	0.03	0.15

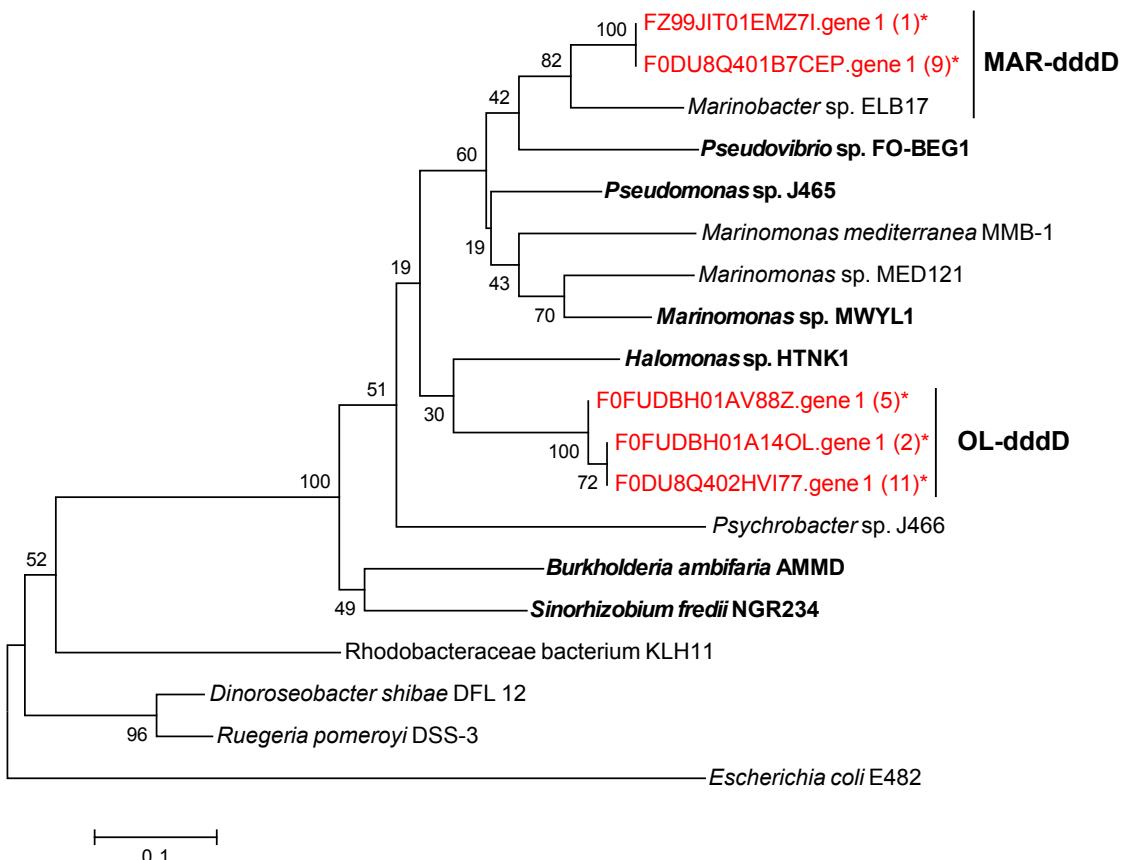


Figure 4.14: Phylogenetic tree of DddD DMSP lyase homologues. *E. coli* carnitine coenzyme A transferase was used as an outgroup. *Dinoroseobacter shibae* DFL 12 and *Ruegeria pomeroyi* DSS-3 homologues are a non-functional outgroup (Todd *et al.*, 2011). The tree was computed from a 75 amino acid region within the conserved amino-terminal class III coenzyme A domain (CaiB) using the neighbour-joining algorithm. Organic Lake sequences from this study are shown in red and marked with an asterisk (*). Numbers in parentheses are counts of sequences that clustered with the Organic Lake homologue shown in the tree with 90% amino acid identity. Sequences with confirmed DMSP lyase activity are shown in bold. Accession numbers from top to bottom are: EBA01716, AEV37420, ACY01992, ADZ91595, EAQ63474, ABR72937, ACV84065, ACY02894, ABI89851, YP_002822700, EEE36156, ABV95365, AAV94987 and EGB36199.

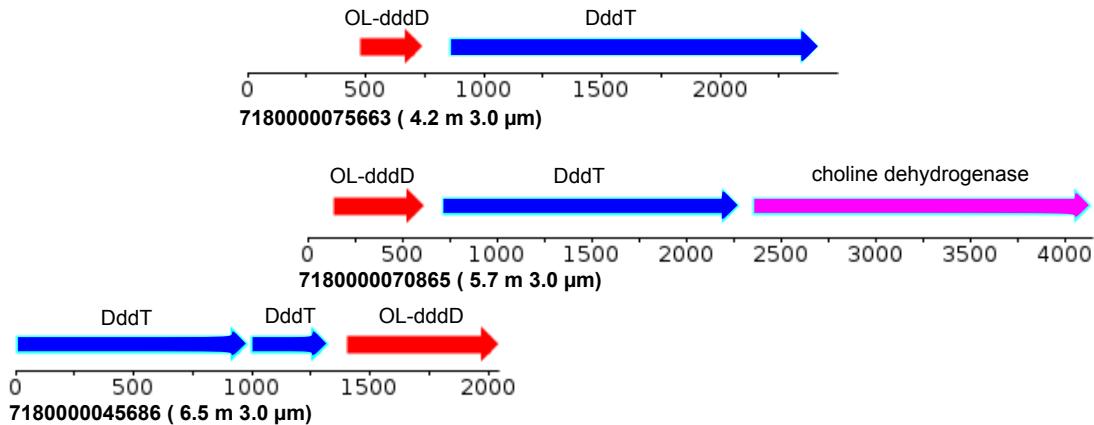


Figure 4.15: Genomic maps of Organic Lake scaffolds containing the OL-dddD homologue. DddT and choline dehydrogenase had best BLAST matches to *Halomonas* sp. HTNK1 (*Gammaproteobacteria*) and *Hoeflea phototrophica* DFL-43 (*Alphaproteobacteria*), respectively. The numbers represent base pairs. The sample depth and filter from which the scaffold was assembled is shown in parentheses beside the scaffold ID.

from Organic Lake *Marinobacter* (Figure 4.14). OLn-dddD did not have a close relative from cultured bacteria making its precise taxonomic origins uncertain. The abundance of OL-dddD on the 3.0 μm fraction suggests it originates from *Alphaproteobacteria*. OL-dddD containing contigs carried genes of mixed *Alpha-* and *Gammaproteobacterial* origin supporting its provenance from one of these classes and consistent with the “pick n’ mix” arrangement of genes found beside sequenced *dddD* regions (Johnston *et al.*, 2008) (Figure 4.15). Adjacent to OL-dddD was *dddT* (Figure 4.15), a betaine, choline, carnitine transporter (BCCT) family protein that likely functions in substrate import, demonstrating OL-dddD forms an operon-like structure, similar to *Halomonas* sp. HTNK1 (Todd *et al.*, 2010).

Two *dddL* groups were detected in Organic Lake: SUL-dddL and MAR-dddL (Figure 4.16). The former includes the *Sulfitobacter* sp. EE-36 *dddL* and the latter the *Marinobacter manganoxydans* MnI7-9 homolog indicating they originate from *Roseobacter* clade and *Gammaproteobacteria*, respectively. *Sulfitobacter* sp. EE-36 has demonstrated DMSP lyase activity and the *dddL* gene alone is sufficient for DMS generation (Curson *et al.*, 2008). These data indicate that the Organic Lake members of the SUL-dddL group perform the same functional role. The MAR-dddL clade appears to be an uncharacterized branch of the *dddL* family. emphdddP was detected as the least abundant of the DMSP lyases (Table 4.6). Phylogenetic analyses showed Organic Lake *dddP* likely originate from *Roseovarius* (Figure 4.17). The Organic Lake sequences formed a clade with the functionally verified *Roseovarius nibinhicens* ISM *dddP* (Todd *et al.*, 2009).

A single type of DMSP demethylase, *dmdA* was identified. It clustered with *Roseobacter* clade *dmdA* (Figure 4.18), corresponding to the marine clade A (Howard *et al.*, 2006), and includes the functionally verified *R. pomeroyi* DSS-3 homolog. These data indicate that the Organic Lake sequences correspond to true DMSP demethylases and not related glycine cleavage T proteins or aminomethyltransferases (Howard *et al.*, 2006). DMSP

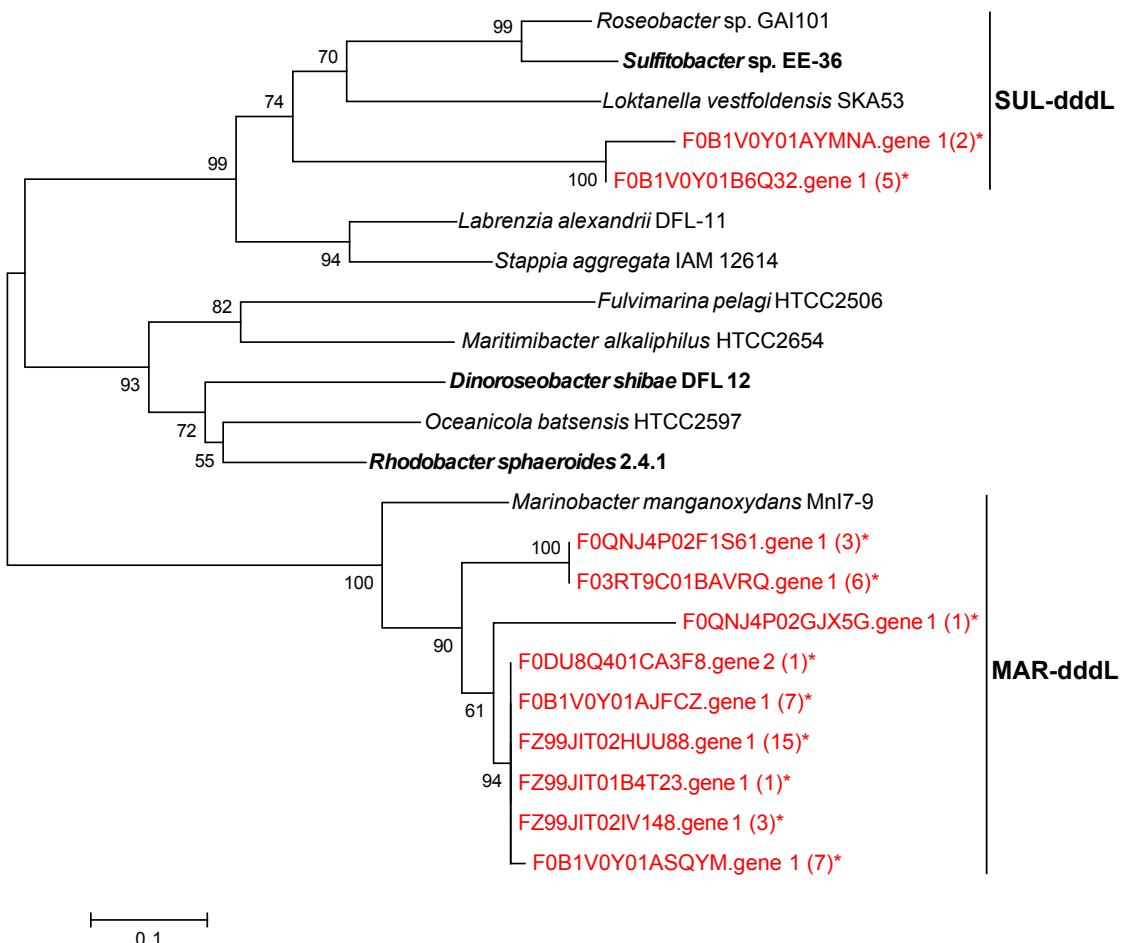


Figure 4.16: Phylogenetic tree of DddL DMSP lyase homologues. The tree was computed from an 84 amino acid N-terminal region using the neighbour-joining algorithm. Organic Lake sequences from this study are shown in red and marked with an asterisk (*). Numbers in parentheses are counts of sequences that clustered with the Organic Lake homologue shown in the tree with 90% amino acid identity. Sequences with confirmed DMSP lyase activity are shown in bold. Accession numbers from top to bottom are: EEB86351, ADK55772, EAQ07081, EEE47811, EAV43167, EAU41122, EAQ10619, ABV95046, EAQ04071, ABA77574 and EHJ04839.

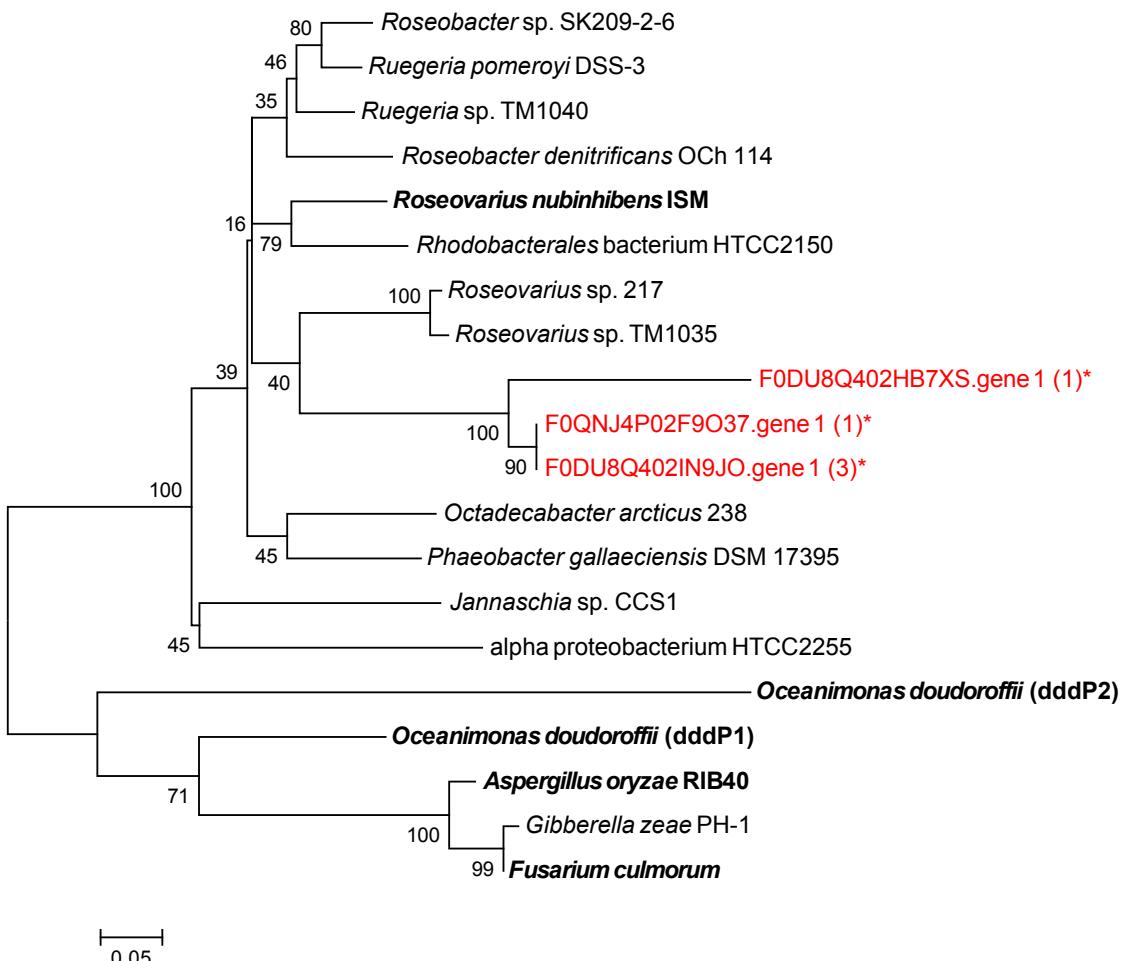


Figure 4.17: Phylogenetic tree of DddP DMSP demethylase homologues. The tree was computed from a 129 amino acid C-terminal region using the neighbour-joining algorithm. Organic Lake sequences from this study are shown in red and marked with an asterisk (*). Numbers in parentheses are counts of sequences that clustered with the Organic Lake homologue shown in the tree with 90% amino acid identity. Sequences with confirmed DMSP lyase activity are shown in bold. Accession numbers from top to bottom are: ZP_01755203, YP_167522, YP_613011, YP_682809, EAP77700, ZP_01741265, ZP_01036399, ZP_01881042, ZP_05063825, AFO91571, YP_509721, ZP_01448542, AEQ39103, AEQ39091, XP_001823911, XP_389272 and ACF19795.

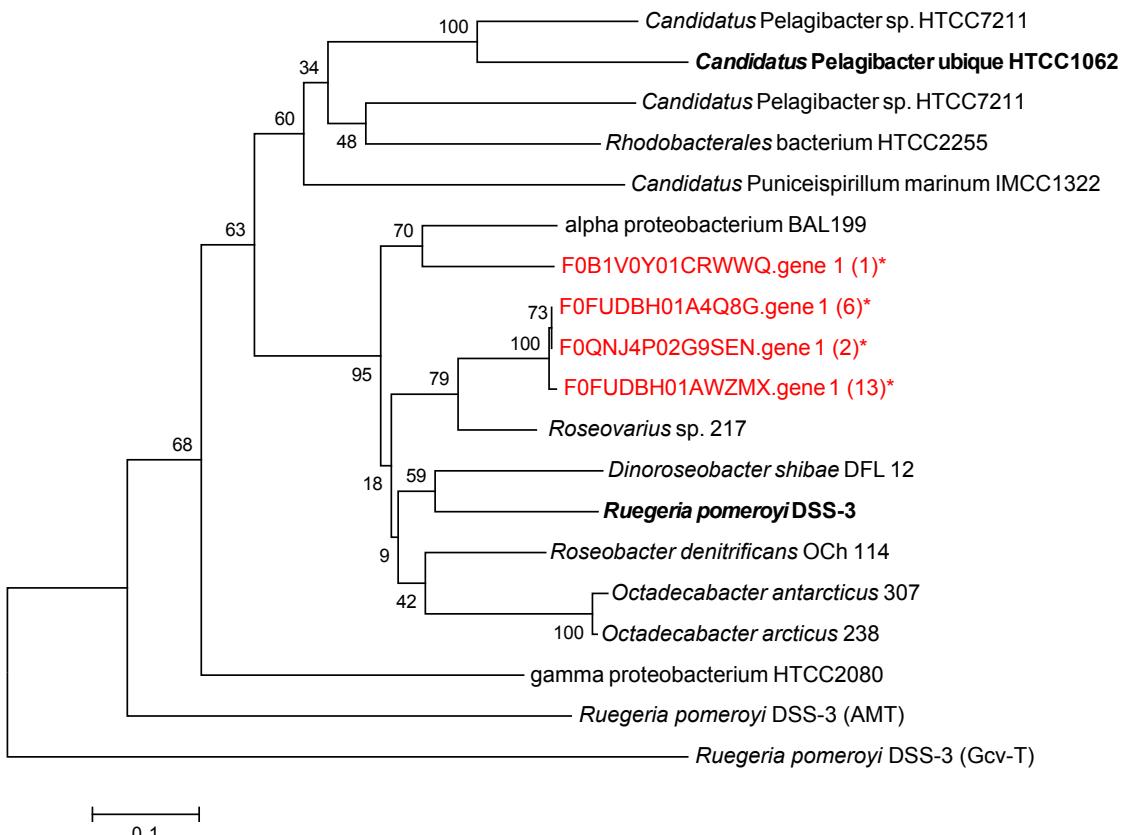


Figure 4.18: Phylogenetic tree of DmdA DMSP demethylase homologues. The tree was computed from a 128 amino acid region using the neighbour-joining algorithm. Organic Lake sequences from this study are shown in red and marked with an asterisk (*). Numbers in parentheses are counts of sequences that clustered with the Organic Lake homologue shown in the tree with 90% amino acid identity. Sequences with confirmed DMSP demethylase activity are shown in bold. Accession numbers from top to bottom are: EDZ60447, YP_265671, EDZ61098, EAU51039, YP_003550401, EDP61332, EAQ26389, ABV94056, AAV94935, AAV95190, EDY79173, EDY89914, EAW42451, AAV94935 and AAV97197.

cleavage appears to be a significant source of DMS in Organic Lake. DMSP likely originates from *Bacillariophyta* or *Dinoflagellida* as Organic Lake *Dunaliella* have been reported not to produce DMSP in culture (Franzmann *et al.*, 1987b). Based on the abundance of marker genes, DMSP cleavage is predicted to occur at highest levels in the deep zone (Figure 4.13) where the DMS concentration has been measured to be highest (Deprez *et al.*, 1986; Franzmann *et al.*, 1987b; Gibson *et al.*, 1991; Roberts and Burton, 1993; Roberts *et al.*, 1993). DMS can also be produced in anoxic environments from the reduction of DMSO, degradation of sulphur containing amino acids, and sulphide methylation (Schäfer *et al.*, 2010). Our data indicate that some DMSO reduction linked to *Firmicutes* could occur, but is not likely a major pathway (Figure 4.13), and the potential for the other DMS yielding processes could not be determined because the enzymes involved in these pathways have not been established. When cultivated, Halomonas isolates from Organic Lake produced DMS from cysteine (Franzmann *et al.*, 1987b) providing some evidence that DMS production from anaerobic degradation of amino acids can occur. Abiotic pathways for anaerobic production of DMS have also been proposed (Roberts *et al.*, 1993).

The potential for DMSP cleavage was more than twice that of DMSP demethylation (Figure 4.13). This is unusual compared to the marine environment or Ace Lake where DMSP demethylation potential is much higher than cleavage (Table 4.6). Previous estimates have similarly shown marine environments to have demethylation potential up to two orders of magnitude higher than cleavage (Howard *et al.*, 2008; Todd *et al.*, 2009, 2011; Reisch *et al.*, 2011). The frequency of DMSP lyase genes *dddD* and *dddL* in Organic Lake exceeded those of all other environments, except Punta Cormorant hypersaline lagoon, where *dddL* abundance was comparable (Table 4.6). This suggests selection in Organic Lake for DMSP cleavage due to functional advantage and/or selection for taxa that carry DMSP lyase genes. There is evidence that high DMSP cleavage potential is adaptive in hypersaline systems, as a high proportion of *ddd* genes were similarly detected in Punta Cormorant hypersaline lagoon and saltern ponds (Raina *et al.*, 2010). Determination of the taxonomic composition of these other hypersaline environments could indicate whether selection is occurring for functional capacity or on a taxonomic level if the taxonomic composition between these systems was significantly different but abundance of DMSP lyase genes were high.

The accumulated DMS in Organic Lake suggests conditions in Organic Lake favor the relatively inefficient lysis pathway, where both sulphur and carbon is lost to the organism performing the DMSP lysis, over the more ‘thrifty’ demethylation pathway. This is particularly pertinent to the *Roseobacter* lineages that can also perform either process. One possibility that has been proposed is that when sulphur is in excess and the organism can easily assimilate alternative sulphur sources, the lysis pathway may be competitive (Johnston *et al.*, 2008). This may be particularly the case in hypersaline systems if higher concentrations of DMSP are being produced as an osmolyte.

4.4.8 Conclusion

Through the use of shotgun metagenomics and size partitioning of samples, we discovered that the Organic Lake system is dominated by heterotrophic bacteria related to *Psychroflexus*, *Marinobacter* and *Roseovarius* with primary production provided largely by chlorophyte algae related to *Dunaliella*. Genetic potential for oxidation of fixed carbon by heterotrophic bacteria occurs greatly in excess of carbon fixation, suggesting possible net carbon loss. However, by linking key metabolic processes to the dominant heterotrophic lineages we uncovered processes that were unusually abundant in Organic Lake that may serve to maximize exploitation of limited resources and minimize loss. Recalcitrant polymeric algal material and particulate matter is likely remineralized by *Psychroflexus* in the upper mixed zone and by *Firmicutes* in the deep zone to provide labile substrates for use by other heterotrophic bacteria. The generalist *Marinobacter* and *Roseovarius* lineages were associated with abundant genes involved in rhodopsin-mediated and AAnP photoheterotrophy; the latter of which was more abundant in Organic Lake than any other system surveyed. Potential for chemolithoheterotrophy, sulphur oxidation and CO oxidation was also high, and along with photoheterotropy, may provide a supplementary energy source if organic carbon becomes limiting.

In addition to being able to describe the functional capacities and potential importance of poorly understood microbial processes occurring in the lake (e.g. photoheterotrophy by *Alphaproteobacteria*), we were able to answer targeted questions about the biology of the unusual lake sulphur chemistry. The low potential for dissimilatory sulphur cycling in the deep zone and relatively stable waters, combined with the generation of DMS from DMSP, facilitate the accumulation of a high level of DMS in the lake. It appears *Marinobacter* and *Roseovarius* play a key role in DMS formation by cleaving DMSP generated by upper mixed zone eucaryal algae. The remarkable abundance of DMSP lyase genes suggests DMSP is a significant carbon source in Organic Lake and the cleavage pathway provides a selective advantage under the unique constraints of the Organic Lake environment.

In view of the minimal capacity for biological fixation of carbon and nitrogen, and yet organic richness, including high levels of DMS, in Organic Lake, we evaluated what input the lake may have received throughout its relatively brief \sim 3,000 year history. The volume of the lake is small (\sim 6 \times 10⁴ m³), and exogenous input may occur from guano deposits in a small penguin rookery nearby the lake, through giant petrel or skua predation and defecation, and/or by decaying animal carcasses such as elephant seals which can weigh on the order of 1 ton and are present near the lake. It is also possible that during isolation from the ocean, the base of the water column in the marine basin that formed the lake may have acted as a sump for organic material. Phytoplankton blooms and benthic mats tend to make coastal marine basins very productive, and organic matter that sediments out of the surface waters will become trapped in the denser, more saline bottom layers (Bird *et al.*, 1991). Retention of captured organic matter in the lake may also have been facilitated by Organic Lake having become highly saline quickly (Bird *et al.*, 1991). Studies in the future that experimentally determine

exogenous input and historical lake dynamics (e.g. stable isotope and biomarker analyses of lake sediment), the role of benthic communities, and metaproteogenomic analyses of interannual community composition and function, will provide improved knowledge of the unusual biogeochemistry of Organic Lake and better enable predictions to be made about how the lake may be affected by ecosystem changes.

4.5 Acknowledgements

This work was supported by the Australian Research Council and the Australian Antarctic Science program. The authors acknowledge assistance from the J. Craig Venter Institute and the Gordon and Betty Moore Foundation. We thank John Bowman for providing unpublished rhodopsin sequence data.

Chapter 5

General discussion, conclusions and future work

Some ideas for this section include

Perspective of Antarctic Lake research from wetlab to molecular age. Summary of all molecular work done by our group to present. Summary of the major achievements of my work.

Needed future work for virophages. Since publication of my work, more virophages have been found. Need to isolate and track them over a season. Determine which OLPV type it infects. Verify OLPV infects pyramimonas. Verify OLV reduces infective particles. Make an exclusion experiment to show that the dynamics change with and without the OLV.

References

- Abell G. C. J. and Bowman J. P. (2005). Ecological and biogeographic relationships of class Flavobacteria in the Southern Ocean. *FEMS microbiology ecology*, 51(2):265–77.
- Abell G. C. J. and Bowman J. P. (2005). Colonization and community dynamics of class Flavobacteria on diatom detritus in experimental mesocosms based on Southern Ocean seawater. *FEMS microbiology ecology*, 53(3):379–391.
- Alekhina I. A., Marie D., Petit J. R., Lukin V. V., Zubkov V. M., and Bulat S. A. (2007). Molecular analysis of bacterial diversity in kerosene-based drilling fluid from the deep ice borehole at Vostok, East Antarctica. *FEMS microbiology ecology*, 59(2):289–299.
- Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215:403–410.
- Balashov S. P., Imasheva E. S., Boichenko V. a., Antón J., Wang J. M., and Lanyi J. K. (2005). Xanthorhodopsin: a proton pump with a light-harvesting carotenoid antenna. *Science (New York, N.Y.)*, 309(5743):2061–4.
- Béjà O., Suzuki M. T., Heidelberg J. F., Nelson W. C., Preston C. M., Hamada T., Eisen J. A., Fraser C. M., and Delong E. F. (2002). Unsuspected diversity among marine aerobic anoxygenic phototrophs. 415:5–8.
- Bell E. M. and Laybourn-Parry J. (2003). MIXOTROPHY IN THE ANTARCTIC PHYTOFLAGELLATE , PYRAMIMONAS GELIDICOLA (CHLOROPHYTA : PRASINOPHYCEAE). *Journal of Phycology*, 649:644–649.
- Bengtsson J., Eriksson K. M., Hartmann M., Wang Z., Shenoy B. D., Grelet G.-A., Abarenkov K., Petri A., Rosenblad M. A., and Nilsson R. H. (2011). Metaxa : a software tool for automated detection and discrimination among ribosomal small sub-unit mitochondria , and chloroplasts in metagenomes and environmental sequencing datasets. *Antonie van Leeuwenhoek*, 100:471–475.
- Bergh O. i., BØ rsheim K. Y., Bratbak G., and Heldal M. (1989). High abundance of viruses found in aquatic environments. *Nature*, 340:467–468.
- Bergmann D. J., Hooper A. B., and Klotz M. G. (2005). Structure and sequence conservation of hao cluster genes of autotrophic ammonia-oxidizing bacteria : evidence for their evolutionary history. *Applied and environmental microbiology*, 71(9):5371–5382.
- Bielewicz S., Bell E. M., Kong W., Friedberg I., Priscu J. C., and Morgan-Kiss R. M. (2011). Protist diversity in a permanently ice-covered Antarctic lake during the polar night transition. *The ISME journal*, 5(9):1559–1564.
- Bird M. I., Chivas A. R., Radnell C. J., and Burton H. R. (1991). Sedimentological and stable-isotope evolution of lakes in the Vestfold Hills, Antarctica. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 84:109–130.

- Bowman J. P., Mccammon S. A., Lewis T., Skerratt J. H., Brown J. L., Nichols D. S., and Mcmeekin T. A. (1998). Psychroflexus torquis gen. nov., sp. nov., a psychrophilic species from Antarctic sea ice, and reclassification of Flavobacterium gondwanense (Dobson et al. 1993) as Psychroflexus gondwanense gen. nov., comb. nov. *Microbiology*, 144:1601–1609.
- Bowman J. P., McCammon S. A., Rea S. M., and McMeekin T. A. (2000). The microbial composition of three limnologically disparate hypersaline Antarctic lakes. *FEMS Microbiology Letters*, 183(1):81–88.
- Bowman J. P., Rea S. M., McCammon S. A., and McMeekin T. A. (2000). Diversity and community structure within anoxic sediment from marine salinity meromictic lakes and a coastal meromictic marine basin, Vestfold Hills, Eastern Antarctica. *Environmental Microbiology*, 2(2):227–237.
- Bronge C. (2004). Hydrographic and climatic changes influencing the proglacial Druzhby drainage system, Vestfold Hills, Antarctica. *Antarctic Science*, 8(4):379–388.
- Budinoff C. R., Loar S. N., LeCleir G. R., Wilhelm S. W., and Buchan A. (2011). A protocol for enumeration of aquatic viruses by epifluorescence microscopy using Anodisc 13 membranes. *BMC microbiology*, 11(1):168.
- Bulat S. A., Alekhina I. A., Blot M., Petit J.-R., Angelis M.de, Wagenbach D., Lipenkov V. Y., Vasilyeva L. P., Wloch D. M., Raynaud D., and Lukin V. V. (2004). DNA signature of thermophilic bacteria from the aged accretion ice of Lake Vostok, Antarctica: implications for searching for life in extreme icy environments. *International Journal of Astrobiology*, 3(1):1–12.
- Burke C. and Burton H. R. (1988). Photosynthetic bacteria in meromictic lakes and stratified fjords of the Vestfold Hills, Antarctica. *Hydrobiologia*, 165:13–23.
- Burton H. R. (1981). Chemistry , physics and evolution of Antarctic saline lakes. *Hydrobiologia*, 82(1):339–362.
- Campbell B. J., Engel A. S., Porter M. L., and Takai K. (2006). The versatile epsilon-proteobacteria: key players in sulphidic habitats. *Nature reviews. Microbiology*, 4(6): 458–468.
- Caporaso J. G., Kuczynski J., Stombaugh J., Bittinger K., Bushman F. D., Costello E. K., Fierer N., Peña A. G., Goodrich J. K., Gordon J. I., Huttley G. A., Kelley S. T., Knights D., Koenig J. E., Ley R. E., Lozupone C. A., McDonald D., Muegge B. D., Pirrung M., Reeder J., Sevinsky J. R., Turnbaugh P. J., Walters W. A., Widmann J., Yatsunenko T., Zaneveld J., and Knight R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Publishing Group*, 7(5):335–336.
- Charlson R. J., Lovelock J. E., Andreae M. O., and Warren S. G. (1987). Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature*, 326:655–661.
- Chouari R., Le Paslier D., Daegelen P., Ginestet P., Weissenbach J., and Sghir A. (2005). Novel predominant archaeal and bacterial groups revealed by molecular analysis of an anaerobic sludge digester. *Environmental microbiology*, 7(8):1104–1115.
- Christner B. C., Mosley-Thompson E., Thompson L. G., and Reeve J. N. (2001). Isolation of bacteria and 16S rDNAs from Lake Vostok accretion ice. *Environmental Microbiology*, 3(9):570–577.

- Clarke K. and RN G. PRIMER V6: User Manual/Tutorial, 2006.
- Cottrell M. T. and Kirchman D. L. (2009). Photoheterotrophic microbes in the Arctic Ocean in summer and winter. *Applied and environmental microbiology*, 75(15):4958–4966.
- Curran M. A., Jones G. B., and Burton H. R. (1998). Spatial distribution of dimethylsulphide and dimethylsulfoniopropionate in the Australiasian sector of the Southern Ocean. *Journal of Geophysical Research*, 103:16677–16689.
- Curson A. R., Rogers R., Todd J. D., Brearley C., and Johnston A. W. (2008). Molecular genetic analysis of a dimethylsulfoniopropionate lyase that liberates the climate-changing gas dimethylsulfide in several marine alpha-proteobacteria and Rhodobacter sphaeroides. *Environmental microbiology*, 10(3):757–767.
- Curson A. R., J , Sullivan M. J., Todd J. D., and Johnston A. W. (2010). Identification of genes for dimethyl sulfide production in bacteria in the gut of Atlantic Herring (*Clupea harengus*). *The ISME journal*, 4(1):144–146.
- Curson A. R., Todd J. D., Sullivan M. J., and Johnston A. W. (2011). Catabolism of dimethylsulphoniopropionate: microorganisms , enzymes and genes. *Nature Reviews Microbiology*, 9:849–859.
- Demergasso C., Escudero L., Casamayor E. O., Chong G., Balagué V., and Pedrós-Alio C. (2008). Novelty and spatio-temporal heterogeneity in the bacterial diversity of hypersaline Lake Tebenquiche (Salar de Atacama). *Extremophiles*, 12(4):491–504.
- Deprez P. P., Franzmann P. D., and Burton H. R. (1986). Determination of reduced sulfur gases in Antarctic lakes and seawater by gas chromatography after solid adsorbent preconcentration. *Journal of Chromatography*, 362:9–21.
- Diemer G. S., Kyle J. E., and Stedman K. M. Counting viruses using polycarbonate Track Etch membrane filters as an alternative to Anodisc membrane filters ., 2012. URL http://www.web.pdx.edu/~kstedman/PCTE_virus_counting_protocol.pdf.
- Dobson S., James S., Franzmann P. D., and Mcmeekin T. A. (1991). A numerical taxonomic study of some pigmented bacteria isolated from Organic Lake, an antarctic hypersaline lake. *Archives of Microbiology*, 156:56–61.
- Edwards K., Rogers D., Wirsen C., and McCollom T. (2003). Isolation and characterization of novel psychrophilic , neutrophilic , Fe-oxidizing, chemolithoautotrophic alpha- and gamma-Proteobacteria from the deep sea. *Applied and environmental microbiology*, 69(5):2906–2913.
- Elshahed M. S., Najar F. Z., Aycock M., Qu C., Roe B. A., and Krumholz L. R. (2005). Metagenomic Analysis of the Microbial Community at Zodletone Spring (Oklahoma): Insights into the Genome of a Member of the Novel Candidate Division OD1. *Applied and environmental microbiology*, 71(11):7598–7602.
- Ferris J. M., Gibson J. A., and Burton H. R. (1991). Evidence of density currents with the potential to promote meromixis in ice-covered saline lakes. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 84:99–107.
- Fofonoff N. and Millard R. J. (1983). Algorithms for computation of fundamental properties of seawater. *UNESCO technical papers in marine science*, 44.

- Franzmann P. D., Burton H. R., and Mcmeekin T. A. (1987). Halomonas subglaciescola, a new species of halotolerant bacteria isolated from Antarctica. *International journal of systematic bacteriology*, 37:27–34.
- Franzmann P. D., Deprez P. P., Burton H. R., and Hoff J.van den (1987). Limnology of Organic Lake, Antarctica, a meromictic lake that contains high concentrations of dimethyl sulfide. *Marine and Freshwater Research*, 38(3):409–417.
- Franzmann P. D., Stackebrandt E., Sanderson K., Volkman J. K., Cameron D., Steven-son P., and Mcmeekin T. A. (1988). Halorubrum lacusprofundi sp. nov., a halophilic bacterium idolated from Deep Lake, Antarctica. *Systematic & Applied Microbiology*, 11:20–27.
- Friedrich C. G., Bardischewsky F., Rother D., Quentmeier A., and Fischer J. (2005). Prokaryotic sulfur oxidation. *Current opinion in microbiology*, 8(3):253–259.
- Fuhrman J. A., Schwalbach M. S., and Stingl U. (2008). Proteorhodopsins: an array of physiological roles? *Nature reviews. Microbiology*, 6:488–494.
- Gärdes A., Kaeppel E., Shehzad A., Seebah S., Teeling H., Yarza P., Glöckner F. O., Grossart H.-P., and Ullrich M. S. (2010). Complete genome sequence of Marinobacter adhaerens type strain (HP15), a diatom-interacting marine microorganism. *Standards in genomic sciences*, 3(2):97–107.
- Gauthier M., Lafay B., Christen R., Fernandez L., Acquaviva M., Bonin P., and Bertrand J.-C. (1992). a New , Extremely Halotolerant , Hydrocarbon-Degrading Marine Bac- terium. *International journal of systematic bacteriology*, (4):568–576.
- Gibson J. A. (1999). The meromictic lakes and stratified marine basins of the Vestfold Hills, East Antarctica. *Antarctic Science*, 11(2):175–192.
- Gibson J. A., Ferris J. M., and Burton H. R. (1990). Temperature density, tempera- ture conductivity and conductivity-density relationships for marine-derived saline lake waters. *ANARE research notes*, 78.
- Gibson J. A., Garrick R., Franzmann P. D., Deprez P. P., and Burton H. R. (1991). Reduced sulfur gases in saline lakes of the Vestfold Hills, Antarctica. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 84(1-4):131–140.
- Gibson J. A., Qiang X. L., Franzmann P. D., Garrick R. C., and Burton H. R. (1994). Volatile fatty and dissolved free amino acids in Organic Lake, Vestfold Hilss, East Antarctica. *Polar Biology*, 14:545–550.
- Glatz R., Lepp P., Ward B. B., and Francis C. (2006). Planktonic microbial community composition across steep physical/chemical gradients in permanently ice-covered Lake Bonney, Antarctica. *Geobiology*, 4:53–67.
- Goberna M., Insam H., and Franke-Whittle I. (2009). Effect of biowaste sludge matu- ration on the diversity of thermophilic bacteria and archaea in an anaerobic reactor. *Applied and environmental microbiology*, 75(8):2566–2572.
- Gómez-Consarnau L., González J. M., Coll-Lladó M., Gourdon P., Pascher T., Neutze R., Pedrós-Alio C., and Pinhassi J. (2007). Light stimulates growth of proteorhodopsin- containing marine Flavobacteria. 445:210–213.

- Gómez-Consarnau L., Akram N., Lindell K., Pedersen A., Neutze R., Milton D. L., González J. M., and Pinhassi J. (2010). Proteorhodopsin phototrophy promotes survival of marine bacteria during starvation. *PLoS Biology*, 8(4):2–11.
- Gordon D., Priscu J. C., and Giovanonni S. J. (2000). Origin and Phylogeny of Microbes Living in Permanent Antarctic Lake Ice. *Microbial ecology*, 39(3):197–202.
- Green D. H., Bowman J. P., Smith E. A., Gutierrez T., and Bolch C. J. (2006). Marinobacter algicola sp. nov., isolated from laboratory cultures of paralytic shellfish toxin-producing dinoflagellates. *International journal of systematic and evolutionary microbiology*, 56:523–527.
- Green W. J., Angle M. P., and Chave K. E. (1988). The geochemistry of Antarctic streams and their role in the evolution of four lakes of the McMurdo Dry Valleys. *Geochimica et Cosmochimica Acta*, 52(5):1265–1274.
- Hahn M. W. (2009). Description of seven candidate species affiliated with the phylum Actinobacteria, representing planktonic freshwater bacteria. *International journal of systematic and evolutionary microbiology*, 59:112–117.
- Hahn M. W., Stadler P., Wu Q. L., and Pöckl M. (2004). The filtration-acclimatization method for isolation of an important fraction of the not readily cultivable bacteria. *Journal of microbiological methods*, 57(3):379–390.
- Hahsler M., Hornik K., and Buchta C. (2007). Getting things in order: an introduction to the R package seriation. *Journal of Statistical Software*, 25(3):1–34.
- Harris J. K., Kelley S. T., and Pace N. R. (2004). New Perspective on Uncultured Bacterial Phylogenetic Division OP11. *Applied and environmental microbiology*, 70 (2):845–849.
- Hodgson D. A. Antarctic Lakes. In Bengtsson L., Herschy R. W., and Fairbridge R. W., editors, *Encyclopedia of Lakes and Reservoirs*, pages 26–31. Springer, Dordrecht, 2012.
- Howard E. C., Henriksen J. R., Buchan A., Reisch C. R., Bürgmann H., Welsh R., Ye W., González J. M., Mace K., Joye S. B., Kiene R. P., Whitman W. B., and Moran M. A. (2006). Bacterial taxa that limit sulfur flux from the ocean. *Science (New York, N.Y.)*, 314(5799):649–652.
- Howard E. C., Sun S., Biers E. J., and Moran M. A. (2008). Abundant and diverse bacteria involved in DMSP degradation in marine surface waters. *Environmental microbiology*, 10(9):2397–2410.
- Huang L.-N., Zhu S., Zhou H., and Qu L.-H. (2005). Molecular phylogenetic diversity of bacteria associated with the leachate of a closed municipal solid waste landfill. *FEMS microbiology letters*, 242(2):297–303.
- Huang Y., Niu B., Gao Y., Fu L., and Li W. (2010). CD-HIT Suite : a web server for clustering and comparing biological sequences. *Bioinformatics*, 26:260–262.
- Hügler M. and Sievert S. M. (2011). Beyond the Calvin Cycle: Autotrophic Carbon Fixation in the Ocean. *Annual Review of Marine Science*, 3(1):261–289.
- Humayoun S. B., Bano N., and Hollibaugh J. T. (2003). Depth distribution of microbial diversity in Mono Lake , a meromictic soda lake in California. 69(2):1030–1042.

- Huu N. B., Denner E. B., Ha D. T., Wanner G., and Stan-Lotter H. (1996). Marinobacter aquaeolei sp.nov., a halophilic bacterium isolated from a Vietnamese oil-producing well. *International journal of systematic bacteriology*, (49):367–375.
- James S., Dobson S., Franzmann P. D., and Mcmeekin T. A. (1990). Halomonas meridiana, a New Species of Extremely Halotolerant Bacteria Isolated from Antarctic Saline Lakes. *Systematic and Applied Microbiology*, 13(3):270–278.
- James S., Burton H. R., Mcmeekin T. A., and Mancuso C. (1994). Seasonal abundance of Halomonas meridiana, Halomonas subglaciescola, Flavobacterium gondwanense and Flavobacterium salegens in four Antarctic lakes. *Antarcti*, 6(3):325–332.
- Johnston A. W., Todd J. D., Sun L., Nikolaïdou-Katsaraïdou N., Curson A. R., and Rogers R. (2008). Molecular diversity of bacterial production of the climate-changing gas, dimethyl sulphide, a molecule that impinges on local and global symbioses. *Journal of experimental botany*, 59(5):1059–1067.
- Johnstone G., Brown D., and Lugg D. (1973). The biology of the Vestfold Hills, Antarctica. *ANARE Scientific Reports*, 123:1–60.
- Kang I., Lee K., Yang S.-J., Choi A., Kang D., Lee Y. K., and Cho J.-C. (2012). Genome sequence of "Candidatus Aquiluna" sp. strain IMCC13023, a marine member of the Actinobacteria isolated from an arctic fjord. *Journal of bacteriology*, 194(13):3550–3551.
- Karr E. A., Sattley W. M., Jung D. O., Madigan M. T., and Achenbach L. A. (2003). Remarkable Diversity of Phototrophic Purple Bacteria in a Permanently Frozen Antarctic Lake. 8(69):4910–4914.
- Karr E. A., Sattley W. M., Rice M. R., Jung D. O., Madigan M. T., and Achenbach L. A. (2005). Diversity and Distribution of Sulfate-Reducing Bacteria in Permanently Frozen Lake Fryxell , McMurdo Dry Valleys , Antarctica. *Applied and Environmental Microbiology*, 71:6353–6359.
- Karr E. A., Ng J. M., Belchik S. M., Matthew W. M., Madigan M. T., Achenbach L. A., and Sattley W. M. (2006). Biodiversity of Methanogenic and Other Archaea in the Permanently Frozen Lake Fryxell , Antarctica. 72(2):1663–1666.
- Kraft B., Strous M., and Tegetmeyer H. E. (2011). Microbial nitrate respiration Genes, enzymes and environmental distribution. *Journal of Biotechnology*, 155:104–117.
- Kurosawa N., Sato S., Kawarabayasi Y., Imura S., and Naganuma T. (2010). Archaeal and bacterial community structures in the anoxic sediment of Antarctic meromictic lake Nurume-Ike. *Polar Science*, 4(2):421–429.
- Labrenz M., Collins M. D., Lawson P. A., Tindall B. J., Schumann P., and Hirsch P. (1999). Roseovarius tolerans gen. nov., sp. nov., a budding bacterium with variable bacteriochlorophyll a production from hypersaline Ekho Lake. *International Journal of Systematic Bacteriology*, 49:137–147.
- Lauro F. M., McDougald D., Thomas T., Williams T. J., Egan S., Rice S., Demaere M. Z., Ting L., Ertan H., Johnson J., Ferriera S., Lapidus A., Anderson I. J., Kyrpides N., Munk A. C., Detter C., Han C. S., Brown M. V., Robb F. T., Kjelleberg S., and Cavicchioli R. (2009). The genomic basis of trophic strategy in marine bacteria. *PNAS*, 106(37):15527–15533.

- Lauro F. M., Demaere M. Z., Yau S., Brown M. V., Ng C., Wilkins D., Raftery M. J., Gibson J. A., Andrews-Pfannkoch C., Lewis M., Hoffman J. M., Thomas T., and Cavicchioli R. (2011). An integrative study of a meromictic lake ecosystem in Antarctica. *The ISME journal*, 5:879–895.
- Law P. (1959). The Vestfold Hills. *ANARE Reports*, 1:1–50.
- Laybourn-Parry J. The microbial loop in Antarctic lakes. In Howard-Willsiams C., Lyons W., and Hawes I., editors, *Ecosystem Dynamics in a Antarctic Ice-Free Landscapes*, pages 231–240. Rotterdam, 1997.
- Laybourn-Parry J., Marshall W. A., and Marchant H. J. (2005). Flagellate nutritional versatility as a key to survival in two contrasting Antarctic saline lakes. *Freshwater Biology*, 50:830–838.
- Ley R. E., Turnbaugh P. J., Klein S., and Gordon J. I. (2006). Human gut microbes associated with obesity. *Nature*, 444:1022–1023.
- López-Bueno A., Tamames J., Velázquez D., Moya A., Quesada A., and Alcamí A. (2009). High diversity of the viral community from an Antarctic lake. *Science*, 326: 858–861.
- Lovelock J. E. and Maggs R. (1972). Atmospheric dimethyl sulphide and the natural sulphur cycle. *Nature*, 237:452–453.
- Man D., Wang W., Sabehi G., Aravind L., Post A. F., Massana R., Spudich E. N., Spudich J. L., and Béjà O. (2003). Diversification and spectral tuning in marine proteorhodopsins. *The EMBO Journal*, 22(8):1725–1731.
- Matsuzaki M., Kubota K., Satoh T., Kunugi M., Ban S., and Imura S. (2006). Dimethyl sulfoxide-respiring bacteria in Suribati Ike, a hypersaline lake, in Antarctica and the marine environment. *Polar Bioscience*, 20:73–81.
- Mikucki J. A. and Priscu J. C. (2007). Bacterial diversity associated with Blood Falls, a subglacial outflow from the Taylor Glacier, Antarctica. *Applied and environmental microbiology*, 73(12):4029–39.
- Mikucki J. A., Pearson A., Johnston D. T., Turchyn A. V., Farquhar J., Schrag D. P., Anbar A. D., Priscu J. C., and Lee P. A. (2009). A contemporary microbially maintained subglacial ferrous "ocean". *Science (New York, N.Y.)*, 324(5925):397–400.
- Miyoshi T., Iwatsuki T., and Naganuma T. (2005). Phylogenetic characterization of 16S rRNA gene clones from deep-groundwater microorganisms that pass through 0.2-micrometer-pore-size filters. *Applied and environmental microbiology*, 71(2):1084–1088.
- Moran M. A. and Miller W. L. (2007). Resourceful heterotrophs make the most of light in the coastal ocean. *Nature reviews. Microbiology*, 5(10):792–800.
- Moran M. A., Belas R., Schell M., González J. M., Sun F., Sun S., Binder B. J., Edmonds J., Ye W., Orcutt B., Howard E. C., Meile C., Palefsky W., Goesmann A., Ren Q., Paulsen I., Ulrich L., Thompson L., Saunders E., and Buchan A. (2007). Ecological genomics of marine Roseobacters. *Applied and environmental microbiology*, 73(14): 4559–69.

- Moran M. A., Reisch C. R., Kiene R. P., and Whitman W. B. (2012). Genomic Insights into Bacterial DMSP Transformations. *Annual Review of Marine Science*, 4(1):523–542.
- Mosier A. C., Murray A. E., and Fritsen C. H. (2007). Microbiota within the perennial ice cover of Lake Vida, Antarctica. *FEMS microbiology ecology*, 59(2):274–288.
- Ng C. *A metaproteomic analysis of microbial communities in Ace Lake, Antarctic*. PhD thesis, University of New South Wales, 2010.
- Ng C., DeMaere M. Z., Williams T. J., Lauro F. M., Raftery M., Gibson J. A., Andrews-Pfannkoch C., Lewis M., Hoffman J. M., Thomas T., and Cavicchioli R. (2010). Metaproteogenomic analysis of a dominant green sulfur bacterium from Ace Lake, Antarctica. *The ISME journal*, 4(8):1002–19.
- Noble R. T. and Fuhrman J. A. (1998). Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquatic Microbial Ecology*, 14:113–118.
- Noguchi H., Park J., and Takagi T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic acids research*, 34(19):5623–5630.
- Olson J., Steppe T., Litaker R., and Paerl H. (1998). N₂-Fixing Microbial Consortia Associated with the Ice Cover of Lake Bonney, Antarctica. *Microbial ecology*, 36(3):231–238.
- Patel A., Noble R. T., Steele J. A., Schwalbach M. S., Hewson I., and Fuhrman J. A. (2007). Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. *Nature protocols*, 2(2):269–276.
- Pearce D. A. (2003). Bacterioplankton community structure in a maritime antarctic oligotrophic lake during a period of holomixis, as determined by denaturing gradient gel electrophoresis (DGGE) and fluorescence in situ hybridization (FISH). *Microbial Ecology*, 46(1):92–105.
- Pearce D. A. (2005). The structure and stability of the bacterioplankton community in Antarctic freshwater lakes, subject to extremely rapid environmental change. *FEMS microbiology ecology*, 53(1):61–72.
- Pearce D. A., Gast C. J., Lawley B., and Ellis-Evans J. C. (2003). Bacterioplankton community diversity in a maritime Antarctic lake, determined by culture-dependent and culture-independent techniques. *FEMS microbiology ecology*, 45(1):59–70.
- Pearce D. A., Gast C. J.van der, Woodward K., and Newsham K. K. (2005). Significant changes in the bacterioplankton community structure of a maritime Antarctic freshwater lake following nutrient enrichment. *Microbiology (Reading, England)*, 151:3237–3248.
- Pickard J., Adamson D. A., and Heath C. W. (1986). The evolution of Watts Lake, Vestfold Hills, East Antarctica, from marine inlet to freshwater lake. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 53:271–288.
- Priscu J. C. (1999). Geomicrobiology of Subglacial Ice Above Lake Vostok, Antarctica. *Science*, 286(5447):2141–2144.
- Proctor L. M. and Fuhrman J. A. (1990). Viral mortality of marine bacteria and cyanobacteria. *Nature*, 343:60–62.

- Purdy K., Nedwell D., and Embley T. (2003). Analysis of the sulfate-reducing bacterial and methanogenic archaeal populations in contrasting Antarctic sediments. *Applied and environmental ...*, 69(6):3181–3191.
- Raina J.-B., Dinsdale E. A., Willis B. L., and Bourne D. G. (2010). Do the organic sulfur compounds DMSP and DMS drive coral microbial associations? *Trends in microbiology*, 18(3):101–108.
- Redfield A., Ketchum B., and Richards F. The influence of organisms on the composition of sea-water. In Hill M., editor, *The Sea*, pages 26–77. 1963. URL <http://www.vliz.be/imis/imis.php?refid=28944>.
- Reisch C. R., Moran M. A., and Whitman W. B. (2011). Bacterial catabolism of dimethylsulfoniopropionate (DMSP). *Frontiers in microbiology*, 2(August):1–12.
- Rivière D., Desvignes V., Pelletier E., Chaussonnerie S., Guermazi S., Weissenbach J., Li T., Camacho P., and Sghir A. (2009). Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge. *The ISME journal*, 3(6):700–714.
- Roberts N. and Burton H. R. (1993). Sampling volatile organics from a meromictic Antarctic lake. *Polar Biology*, 13:359–361.
- Roberts N., Burton H. R., and Pitson G. (1993). Volatile organic compounds from Organic Lake, an Antarctic, hypersaline, meromictic lake. *Antarctic Science*, 5(4):361–366.
- Röske K., Sachse R., Scheerer C., and Röske I. (2012). Microbial diversity and composition of the sediment in the drinking water reservoir Saidenbach (Saxonia, Germany). *Systematic and applied microbiology*, 35(1):35–44.
- Rusch D. B., Halpern A. L., Sutton G., Heidelberg K. B., Williamson S. J., Yooseph S., Wu D., Eisen J. A., Hoffman J. M., Remington K., Beeson K. Y., Tran B., Smith H., Baden-Tillson H., Stewart C., Thorpe J., Freeman J., Andrews-Pfannkoch C., Venter J. E., Li K., Kravitz S., Heidelberg J. F., Utterback T., Rogers Y.-H., Falcón L. I., Souza V., Bonilla-Rosso G., Eguiarte L. E., Karl D. M., Sathyendranath S., Platt T., Bermingham E., Gallardo V., Tamayo-Castillo G., Ferrari M. R., Strausberg R. L., Nealson K., Friedman R., Frazier M., and Venter J. C. (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS biology*, 5(3):e77.
- Samsudin A. A., Evans P. N., Wright A.-D. G., and Al Jassim R. (2011). Molecular diversity of the foregut bacteria community in the dromedary camel (*Camelus dromedarius*). *Environmental microbiology*, 13(11):3024–3035.
- Schäfer H., Myronova N., and Boden R. (2010). Microbial degradation of dimethylsulphide and related C1-sulphur compounds: organisms and pathways controlling fluxes of sulphur in the biosphere. *Journal of experimental botany*, 61(2):315–334.
- Schiaffino M. R., Unrein F., Gasol J. M., Farias M. E., Estevez C., Balagué V., and Izaguirre I. (2009). Comparative analysis of bacterioplankton assemblages from maritime Antarctic freshwater lakes with contrasting trophic status. *Polar Biology*, 32(6):923–936.
- Schmidtova J., Hallam S. J., and Baldwin S. A. (2009). Phylogenetic diversity of transition and anoxic zone bacterial communities within a near-shore anoxic basin: Nitinat Lake. *Environmental microbiology*, 11(12):3233–3251.

- Sharma A. K., Zhaxybayeva O., Papke R. T., and Doolittle W. F. (2008). Actinorhodopsins: proteorhodopsin-like gene sequences found predominantly in non-marine environments. *Environmental microbiology*, 10(4):1039–1056.
- Sharma A. K., Sommerfeld K., Bullerjahn G. S., Matteson A. R., Wilhelm S. W., Jezbera J., Brandt U., Doolittle W. F., and Hahn M. W. (2009). Actinorhodopsin genes discovered in diverse freshwater habitats and among cultivated freshwater Actinobacteria. *The ISME journal*, 3(6):726–737.
- Sieger M. J., Ellis-Evans J. C., Tranter M., Mayer C., Petit J.-R., Salamatian A., and Priscu J. C. (2001). Physical, chemical and biological processes in Lake Vostok and other Antarctic subglacial lakes. *Nature*, 414(6864):603–609.
- Tajima K., Aminov R. I., Nagamine T., Ogata K., Nakamura M., Matsui H., and Benno Y. (1999). Rumen bacterial diversity as determined by sequence analysis of 16S rDNA libraries. *FEMS Microbiology Ecology*, 29(2):159–169.
- Tamura K., Peterson D., Peterson N., Stecher G., Nei M., and Kumar S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, 28(10):2731–2739.
- Tang Y.-Q., Ji P., Hayashi J., Koike Y., Wu X.-L., and Kida K. (2011). Characteristic microbial community of a dry thermophilic methanogenic digester: its long-term stability and change with feeding. *Applied microbiology and biotechnology*, 91(5):1447–1461.
- Thomsen H. A. (2007). Ultrastructural studies of the flagellate and cyst stages of *Pseudopedinella tricostata* (Pedinellales, Chrysophyceae). *British phycological journal*, pages 37–41.
- Todd J. D., Rogers R., Li Y. G., Wexler M., Bond P. L., Sun L., Curson A. R., Malin G., Steinke M., and Johnston A. W. (2007). Structural and regulatory genes required to make the gas dimethyl sulfide in bacteria. *Science (New York, N.Y.)*, 315(5812):666–669.
- Todd J. D., Curson A. R., Dupont C. L., Nicholson P., and Johnston A. W. (2009). The dddP gene, encoding a novel enzyme that converts dimethylsulfoniopropionate into dimethyl sulfide, is widespread in ocean metagenomes and marine bacteria and also occurs in some Ascomycete fungi. *Environmental microbiology*, 11(6):1376–1385.
- Todd J. D., Curson A. R., Nikolaidou-Katsaraidou N., Brearley C. A., Watmough N. J., Chan Y., Page P. C., Sun L., and Johnston A. W. (2010). Molecular dissection of bacterial acrylate catabolism - unexpected links with dimethylsulfoniopropionate catabolism and dimethyl sulfide production. *Environmental microbiology*, 12(2):327–343.
- Todd J. D., Curson A. R., Kirkwood M., Sullivan M. J., Green R. T., and Johnston A. W. (2011). DddQ, a novel, cupin-containing, dimethylsulfoniopropionate lyase in marine roseobacters and in uncultured marine bacteria. *Environmental microbiology*, 13(2):427–438.
- Todd J. D., Kirkwood M., Newton-Payne S., and Johnston A. W. (2012). DddW , a third DMSP lyase in a model Roseobacter marine bacterium , Ruegeria pomeroyi DSS-3. *The ISME Journal*, 6(1):223–226.

Torrice M. Viral ecology hit by filter shortage, 2009. URL <http://news.sciencemag.org/scienceinsider/2009/10/viral-ecology-r.html>.

Unrein F., Izaguirre I., Massana R., Balagué V., and Gasol J. M. (2005). Nanoplankton assemblages in maritime Antarctic lakes : characterisation and molecular fingerprinting comparison. *Aquatic Microbial Ecology*, 40:269–282.

Varin T., Lovejoy C., Jungblut A. D., Vincent W. F., and Corbeil J. (2012). Metagenomic analysis of stress genes in microbial mat communities from Antarctica and the High Arctic. *Applied and environmental microbiology*, 78(2):549–59.

Venter J. C., Remington K., Heidelberg J. F., Halpern A. L., Rusch D., Eisen J. A., Wu D., Paulsen I., Nelson K. E., Nelson W., Fouts D. E., Levy S., Knap A. H., Lomas M. W., Nealson K., White O., Peterson J., Hoffman J., Parsons R., Baden-Tillson H., Pfannkoch C., Rogers Y.-H., and Smith H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, N.Y.)*, 304(5667):66–74.

Villaescusa J. A., Casamayor E. O., Rochera C., Velázquez D., Álvaro C., Quesada A., and Camacho A. (2010). A close link between bacterial community composition and environmental heterogeneity in maritime Antarctic lakes. *International ...*, (126): 67–77.

Voytek M. A., Priscu J. C., and Ward B. B. (1999). The distribution and relative abundance of ammonia-oxidizing bacteria in lakes of the McMurdo Dry Valley, Antarctica. *Hydrobiologia*, 401:113–130.

Wagner-Döbler I. and Biebl H. (2006). Environmental biology of the marine Roseobacter lineage. *Annual review of microbiology*, 60:255–280.

Wang Q., Garrity G. M., Tiedje J. M., and Cole J. R. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16):5261–5267.

Ward B. B. and Priscu J. C. (1997). Detection and characterization of denitrifying bacteria from a permanently ice-covered Antarctic lake. *Hydrobiologia*, (347):57–68.

Ward B. B., Granger J., Maldonado M., Casciotti K., Harris S., and Wells M. (2005). Denitrification in the hypolimnion of permanently ice-covered Lake Bonney, Antarctica. *Aquatic ...*, 38(2000):295–307.

Wilkins D., Lauro F. M., Williams T. J., Demaere M. Z., Brown M. V., Hoffman J. M., Andrews-Pfannkoch C., McQuaid J. B., Riddle M. J., Rintoul S. R., and Cavicchioli R. (2012). Biogeographic partitioning of Southern Ocean microorganisms revealed by metagenomics. *Environmental microbiology*.

Williams T. J., Wilkins D., Long E., Evans F., DeMaere M. Z., Raftery M. J., and Cavicchioli R. (2012). The role of planktonic Flavobacteria in processing algal organic matter in coastal East Antarctica revealed using metagenomics and metaproteomics. *Environmental Microbiology*.

Xie C., Mao X., Huang J., Ding Y., Wu J., Dong S., Kong L., Gao G., Li C., and Wei L. (2011). KOBAS 2.0: a web server for the annotation and identification of enriched pathways and diseases. *Nucleic acids research*, 39:W316–322.

- Xing P., Hahn M. W., and Wu Q. L. (2009). Low taxon richness of bacterioplankton in high-altitude lakes of the Eastern Tibetan Plateau, with a predominance of Bacteroidetes and Synechococcus spp. *Applied and environmental microbiology*, 75(22): 7017–25.
- Yamamoto M. and Takai K. (2011). Sulfur metabolisms in epsilon- and gamma-Proteobacteria in deep-sea hydrothermal fields. *Frontiers in microbiology*, 2:1–8.
- Yamane K., Hattori Y., Ohtagaki H., and Fujiwara K. (2011). Microbial diversity with dominance of 16S rRNA gene sequences with high GC contents at 74 and 98 °C subsurface crude oil deposits in Japan. *FEMS microbiology ecology*, 76(2):220–235.
- Yanagibayashi M., Nogi Y., Li L., and Kato C. (1999). Changes in the microbial community in Japan Trench sediment from a depth of 6292 m during cultivation without decompression. *FEMS microbiology letters*, 170(1):271–279.
- Yau S., Lauro F. M., Demaere M. Z., Brown M. V., Thomas T., Raftery M. J., Andrews-Pfannkoch C., Lewis M., Hoffman J. M., Gibson J. A., and Cavicchioli R. (2011). Virophage control of antarctic algal host virus dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 108:6163–6168.
- Yilmaz P., Iversen M. H., Hankeln W., Kottmann R., Quast C., and Glöckner F. O. (2012). Ecological structuring of bacterial and archaeal taxa in surface ocean waters. *FEMS microbiology ecology*, 81(2):373–385.
- Zwart D., Bird M., Stone J., and Lambeck K. (1998). Holocene sea-level change and ice-sheet history in the Vestfold Hills, East Antarctica. *Earth and Planetary Science Letters*, 155(1-2):131–145.

Chapter 6

Appendices

Table 6.1: Peptide data for Organic Lake metaproteomic analysis. ^aProteins that have some shared peptides; ^b162322406 and 162276024 are protein homologues; ^cA group of proteins containing similar peptides that could not be differentiated by the mass spectral analysis. Only one gene number of that groups is displayed.

Gene ID	Peptide sequences	
162322530 ^a	R.AIDECLWAVSSLSPSSSADV.K K.ALGAQPFNYTDAVDALPNSIK.A R.EGTYFDQVQPFQHHTR.Y R.HSNFAMESIEQTTFNGQADFGR.R K.HYGDWMQIWCQLTLDK.N R.IDNATLQLVLSNATVEGTNTAK.V K.INLDR.A	R.LNFNHPCK.E K.LQLNGQDR.F R.NGDLAYR.T R.NYNVLR.I R.QVCAPR.N R.RVNCTISR.N R.VNCTISR.N
162322348	K.GNVDVYQENK.L	K.IESDAEPSWVR.G
162322406 ^b	R.QNQSCGGGVNVQNGTHVNR.T R.TAFHLDGGLSR.Q	K.TNDGTLVGK.S K.YVSESSSTYTR.F
162313481	K.ITTIPENIGQLVK.I	R.SNLQGVTEEQLMSNK.I
162276060	K.TPTGLEFSLTGR.A	R.VNHTDACSTGNK.E
162300260	R.VDIEGGTPFFLK.E	K.YTFQPSELNTYFSK.E
162276024 ^b	K.LGGGISSR.S R.SEVGFQSTMVGSDVAMQR.K K.NINLLSAGANYGINTVGSSLR.N	R.TSLHMGDVLSR.K R.NPNLQIR.S K.YENGSWNTLGQLIR.G
162275992	K.NDNITLTDK.Q	R.LTVNNSIISK.E
162300108	K.NVVINSEGTIIISAVNNK.G K.QDVITDQTNLNVGR.L	
162319393 ^a	R.AIDECLWAVNTLSPDSSSDVK.V K.ALGAQPFNYTDAIDALPNSVK.A R.EGTYFDQVQPFQHHTR.S R.IDNATLQLVLSNATVEGTNTAK.V R.IMSGMGLAYSN K.INLDR.A R.LNFNHPCK.E	K.LQLNGQDR.F R.NGDLAYR.T R.NYNVLR.I R.QVCAPR.N R.RVNCTISR.N R.VNCTISR.N
162300134 ^c	K.ATAGDTHLGGEDFDNR.M R.IINEPTAAAIAYGLDK.K	R.VEIIANDQGNR.T
162286324 ^c	K.DVPLVANFSAK.F OLV9 K.AGLLSEMDAYSLYQMSR.R K.ELVLSFSSGVK.F K.FGTQASTLFLK.D R.GSSATMSGLLTK.S K.GVASAVESAIGGAK.T K.HIQTTPSMVDK.Y R.ITSDVQVAVK.D K.IYLVVVRPQYR.S	K.MKLENTVEK.M R.NGSQQTWNEFR.G K.NILPYDEFVAYK.T F.NVNPSENTLVDR.N K.SEVLEAK.E K.VSVQSADILNVITK.Q R.YISLHPSQYAK.L K.YTSLGSHIVIDPVR.D
OLV8	K.AGTPIPGVVIYEPSYPR.W	K.TLPVFIPTIK.Y

Continued on next page

Table 6.1 – *Continued from previous page*

Gene ID	Peptide sequences
K.DIGTDMPYFIFDK.D	K.TNGTTPPR.F
K.GGYADYR.S	K.YSEDDTNESIR.N
K.TLLEFGQSK.D	K.QAFIGLQK.T

Table 6.2: Proteins identified in the Ace Lake 5 m sample 0.1 µm size-fraction proteome. (*) Protein group identification: proteins that contain similar peptides that could not be differentiated by the mass spectral analysis were grouped. Only one gene number of that group is displayed. (a-z, aa-pp) Protein ambiguity groups: proteins that have some shared peptides with one or more other proteins from the same sample depth are marked with the same letters.

Gene ID	NSA	COG ID	KO ID	KEGG locus	5 m - COG annotated proteins	
					COG description	KEGG description
167852195f	0.02530	COG1653	K10232	AAur_0459	sugar-binding periplasmic proteins/domains : putative alpha-glucosides-binding ABC transporter (AglE)	
167782381*	0.01724	COG1879	K02058	Ping_2790	periplasmic sugar-binding proteins : bifunctional carbohydrate binding and transport protein	
167813321	0.01388	COG1629		GFO_2756	outer membrane receptor proteins, mostly Fe transport : TonB-dependent outer membrane receptor	
167754347	0.01044	COG1879	K02058	CMM_0792	periplasmic sugar-binding proteins : putative sugar ABC transporter, solute-binding protein	
167701754a	0.00967	COG0834	K09969	SAR11_0953	ABC-type amino acid transport system, periplasmic component : yhdW	
167792775	0.00630	COG1879	K10552	SMc02171	periplasmic sugar-binding proteins : fructose transport system substrate-binding protein	
167932252d	0.00537	COG0715	K02051	SAR11_0807	ABC-type nitrate/sulfonate/taurine/bicarbonate transport systems, periplasmic components	
167759671	0.00493	COG0605		NEMVE_v1g231554	superoxide dismutase	
167751919h	0.00468	COG3740		ROP_69760	phage head maturation protease	
167907426	0.00438	COG1638		SAR11_0266	dicarboxylate-binding periplasmic protein : TRAP dicarboxylate transporter - DctP subunit (mannitol/chloroaromatic compounds)	
167711086	0.00425	COG0834	K02030	TM1040_0294	ABC-type amino acid transport system, periplasmic component : lysine-arginine-ornithine-binding periplasmic protein	
167819184	0.00389	COG2113	K02002	SAR11_1302	ABC-type proline/glycine betaine transport systems, periplasmic components : opuAC	
167680030	0.00346	COG0683	K01999	AAur_1271	ABC-type branched-chain amino acid transport systems, periplasmic component : braC	
167865828b	0.00338	COG0834	K09969	HCH_05807	ABC-type amino acid transport system, periplasmic component	
167684228c	0.00331	COG0459	K04077	SAR11_0162	chaperonin GroEL (HSP60 family)	
167868594d	0.00311	COG0715	K02051	SAR11_0807	ABC-type nitrate/sulfonate/taurine/bicarbonate transport systems, periplasmic components	
167785199c	0.00309	COG0459	K04077	CHU_1828	chaperonin GroEL (HSP60 family)	
167819050	0.00304	COG2113	K02001	Plav_1066	ABC-type proline/glycine betaine transport systems, periplasmic components	
167867034	0.00284	COG0687	K02055	SAR11_1336	spermidine/putrescine-binding periplasmic protein : potD;	
167700934	0.00277	COG0450		SPO3383	peroxiredoxin : thiol-specific antioxidant protein	
167816084a	0.00253	COG0834	K09969	SAR11_0953	ABC-type amino acid transport system, periplasmic component : yhdW	
167714114	0.00179	COG0687	K02055	SAR11_1336	spermidine/putrescine-binding periplasmic protein : potD	
167712994b	0.00175	COG0834	K09969	HCH_05807	ABC-type amino acid transport system, periplasmic component	
167824568	0.00164	COG3181		Dshi_2450	uncharacterized BCR : hypothetical protein	

Continued on next page

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167925495	0.00159	COG1653	K02027	Krad_1380	sugar-binding periplasmic proteins/domains
167695410a	0.00155	COG0834	K09969	SAR11_0953	ABC-type amino acid transport system, periplasmic component : yhdW
167695984*	0.00138	COG1879			periplasmic sugar-binding proteins
167703404	0.00134	COG1012	K00128	AAur_pTC20196	NAD-dependent aldehyde dehydrogenases
167718230	0.00125	COG0683	K01999	AAur_1271	ABC-type branched-chain amino acid transport systems, periplasmic component : braC
167735996	0.00103	COG0591		SAR11_0316	Na+/proline, Na+/panthothenate symporters and related permeases : yjcG
167739054	0.00101	COG1028	K00059	SH0230	dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases) : 3-oxoacyl-[acyl-carrier protein] reductase
167701096	0.00100	COG0776		KRH_03630	bacterial nucleoid DNA-binding protein : HU_IHF family transcriptional regulator
167817334	0.00098	COG0715		FRAAL1422	ABC-type nitrate/sulfonate/taurine/bicarbonate transport systems, periplasmic components
167703266c	0.00095	COG0459	K04077	Noca_3982	chaperonin GroEL (HSP60 family)
167768609	0.00095	COG3181		RD1_2202	uncharacterized BCR
167868532	0.00093	COG3181	K07795	HCH_01639	uncharacterized BCR : putative tricarboxylic transport membrane protein
167865516	0.00088	COG0747		CMM_2185	ABC-type dipeptide/oligopeptide/nickel transport systems, periplasmic components
167911715	0.00083	COG0776	K03530	Sala_0799	bacterial nucleoid DNA-binding protein : DNA-binding protein HU-beta
167736316	0.00082	COG0174	K01915	SAR11_0747	glutamine synthase : glnA
167916441	0.00079	COG1629		BF2044	outer membrane receptor proteins, mostly Fe transport : putative TonB-dependent outer membrane receptor protein
167920571	0.00078	COG0776	K03530	SAR11_0817	bacterial nucleoid DNA-binding protein : hupA
167703332	0.00066	COG1732	K05845	Strop_1633	periplasmic glycine betaine/choline-binding (lipo)protein of an ABC-type transport system (osmoprotectant binding protein)
167662373	0.00063	COG0834		Pden_1025	ABC-type amino acid transport system, periplasmic component : extracellular solute-binding protein, family 3
167890974	0.00062	COG1878		nfa12380	uncharacterized ACR, predicted metal-dependent hydrolases
167824660	0.00061	COG0683	K01999	SAR11_1361	ABC-type branched-chain amino acid transport systems, periplasmic component : livJ2; Leu/Ile/Val-binding protein precursor
167886240	0.00061	COG0335	K02884	CHU_0120	rplS; 50S ribosomal protein L19
167921445	0.00058	COG0811		GFO_0088	biopolymer transport proteins : exbB; ExbB-like MotA/TolQ/ExbB family
167776275ee	0.00055	COG3740	K06904	BL0376	phage head maturation protease
167659892*ee	0.00055	COG3740		ROP_69760	phage head maturation protease
167786475	0.00054	COG0098	K02988	Fjoh_0380	rpsE; 30S ribosomal protein S5
167693676*	0.00054	COG0776		Arth_3916	bacterial nucleoid DNA-binding protein
167818330	0.00048	COG0683		Rxyl_0363	ABC-type branched-chain amino acid transport systems, periplasmic component : extracellular ligand-binding receptor

Continued on next page

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167739596	0.00044	COG0545	K03772	BDI_2705	FKBP-type peptidyl-prolyl cis-trans isomerases 1
167808311c	0.00044	COG0459	K04077	SAR11_0162	chaperonin GroEL (HSP60 family)
167706214	0.00044	COG0683		Tfu_1779	ABC-type branched-chain amino acid transport systems, periplasmic component
167866918	0.00044	COG2885	K03640	SAR11_0598	outer membrane protein and related peptidoglycan-associated (lipo)proteins : ompA; OmpA family
167807477	0.00042	COG0834		MSMEG_5368	ABC-type amino acid transport system, periplasmic component : ehuB; ectoine/hydroxyectoine ABC transporter solute-binding protein
167881416e	0.00040	COG0050	K02358	CHU_3175	GTPases - translation elongation factors : tufB, tuf
167765645f	0.00034	COG1653	K10232	Sare_3967	sugar-binding periplasmic proteins/domains
167730910	0.00033	COG3181		Dshi_2450	uncharacterized BCR
167725574	0.00032	COG0450	K03386	CHU_2724	peroxiredoxin : ahpC; alkyl hydroperoxide reductase, subunit C
167768817	0.00032	COG0740	K00288	CHU_1706	protease subunit of ATP-dependent Clp proteases : methylenetetrahydrofolate dehydrogenase (NADP+)
167886236	0.00031	COG0228	K02959	CHU_0117	rpsP; 30S ribosomal protein S16
167907528	0.00031	COG0591		SAR11_0316	Na+/proline, Na+/panthothenate symporters and related permeases : yjcG
167868396	0.00029	COG2358		PBPRA0389	predicted periplasmic binding protein : putative immunogenic protein
167718328	0.00027	COG1744	K07335	AAur_1253	surface lipoprotein : basic membrane protein A and related proteins
167769503c	0.00027	COG0459		CMS_2756	chaperonin GroEL (HSP60 family)
167818958	0.00026	COG1638		TM1040_0356	dicarboxylate-binding periplasmic protein : TRAP dicarboxylate transporter - DctP subunit
167702878	0.00025	COG1879		Krad_1186	periplasmic sugar-binding proteins : periplasmic binding protein/LacI transcriptional regulator
167665756	0.00025	COG0091	K02890	GFO_2834	rplV; 50S ribosomal protein L22
167730894	0.00023	COG1638		RD1_2185	dicarboxylate-binding periplasmic protein : dctP; C4-dicarboxylate-binding periplasmic protein, putative
167680092	0.00022	COG0094	K02931	Lxx20210	rplE; 50S ribosomal protein L5
167868548	0.00020	COG0834	K10018	SAR11_1210	ABC-type amino acid transport system, periplasmic component : octopine/nopaline transport system substrate-binding protein
167892279	0.00019	COG0834	K02030	Veis_2153	ABC-type amino acid transport system, periplasmic component
167817276	0.00018	COG0347	K04751	Acel_1565	nitrogen regulatory protein PII
167862242	0.00018	COG0087	K02906	BT_2727	rplC; 50S ribosomal protein L3
167933288	0.00018	COG1638		Dshi_3326	dicarboxylate-binding periplasmic protein : TRAP dicarboxylate transporter-DctP subunit
167713980*	0.00016	COG0330	K04088	SAR11_0008	membrane protease subunits, stomatin/prohibitin homologs : hflK
167867886	0.00016	COG3181		Csal_1767	uncharacterized BCR : uncharacterized protein UPF0065
167809873	0.00016	COG0539	K02945	CHU_1951	rpsA; 30S ribosomal protein S1
167713982	0.00015	COG0330	K04087	SAR11_0007	membrane protease subunits, stomatin/prohibitin homologs : hflC

Continued on next page

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167822210	0.00015	COG0776		SCO2950	bacterial nucleoid DNA-binding protein : hup, SCE59.09c; DNA-binding protein Hu (hs1)
167820450*	0.00015	COG1192		tlr0963	ATPases involved in chromosome partitioning : probable cell division inhibitor minD
167820614g	0.00015	COG0174	K01915	Krad_3291	glutamine synthase
167714092	0.00014	COG0683	K01999	SAR11_1346	ABC-type branched-chain amino acid transport systems, periplasmic component : livJ
167817130	0.00014	COG0711	K02109	SCO5369	F0F1-type ATP synthase b subunit
167818902	0.00014	COG1638		SAR11_0864	dicarboxylate-binding periplasmic protein
167866078	0.00013	COG0605	K00518	Arth_2086	superoxide dismutase
167865698	0.00013	COG0740	K01358	AAur_2381	protease subunit of ATP-dependent Clp proteases
167817852	0.00013	COG0683		Noca_3017	ABC-type branched-chain amino acid transport systems, periplasmic component : extracellular ligand -binding receptor
167714042	0.00012	COG0683	K01999	SAR11_1361	ABC-type branched-chain amino acid transport systems, periplasmic component : livJ2; Leu/Ile/Val-binding protein precursor
167821000	0.00012	COG1732		MSMEG_2924	periplasmic glycine betaine/choline-binding (lipo)protein of an ABC-type transport system (osmoprotectant binding protein) : permease binding-protein component
167668848g	0.00011	COG0174	K01915	CMM_1636	glutamine synthase : glnA1
167748683*	0.00010	COG0834		PFL_3548	ABC-type amino acid transport system, periplasmic component
167718146	0.00010	COG0088	K02926	Lxx20320	rplD; 50S ribosomal protein L4
167862420	0.00010	COG1629		FP0112	outer membrane receptor proteins, mostly Fe transport : probable TonB-dependent outer membrane receptor precursor
167696080*	0.00010	COG1344	K02406	Csac_1680	flagellin and related hook-associated proteins
167735512	0.00010	COG0803	K09815	Smed_1697	ABC-type Mn/Zn transport system, periplasmic Mn/Zn-binding (lipo)protein (surface adhesin A)
167719882	0.00009	COG0096	K02994	SAR11_1103	rpsH; 30S ribosomal protein S8
167660431h	0.00009	COG3740	K06904	BL0376	phage head maturation protease
167718250	0.00009	COG2213	K02799	GK1948	phosphotransferase system, mannitol-specific IIBC component
167719862*	0.00008	COG0185	K02965	SAR11_1113	rpsS; 30S ribosomal protein S19
167702806	0.00008	COG0081	K02863	KRH_05860	rplA; 50S ribosomal protein L1
167719824*e	0.00008	COG0050	K02358	SAR11_1130	GTPases - translation elongation factors : tufB, tuf
167817466	0.00008	COG0404		mll1258	glycine cleavage system T protein (aminomethyltransferase) : sarcosine dehydrogenase
167868614	0.00008	COG2113	K02002	SAR11_0797	ABC-type proline/glycine betaine transport systems, periplasmic components : proX
167933120	0.00007	COG0803	K09815	Atu1521	ABC-type Mn/Zn transport system, periplasmic Mn/Zn-binding (lipo)protein (surface adhesin A) : znuA
167718168	0.00007	COG0094		CMS_0295	50S ribosomal protein L5
167868724	0.00007	COG0396	K09013	SAR11_0740	iron-regulated ABC transporter ATPase subunit SufC

Continued on next page

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167718156	0.00007	COG0092	K02982	Krad_0694	ribosomal protein S3
167719956	0.00006	COG0834	K02030	SAR11_1068	ABC-type amino acid transport system, periplasmic component : pheC; cyclohexadienyl dehydratase
167730882	0.00006	COG0004		SAR11_0818	ammonia permeases : amtB; ammonium transporter
167718138e	0.00006	COG0050	K02358	Tfu_2648	GTPases - translation elongation factors: tuf
167718052	0.00006	COG1653	K02027	Krad_3469	sugar-binding periplasmic proteins/domains
167700960*	0.00006	COG3794		SMa1243	plastocyanin : Azu1 pseudoazurin (blue copper protein)
167868482	0.00005	COG0715			ABC-type nitrate/sulfonate/taurine/bicarbonate transport systems, periplasmic components
167868656	0.00005	COG0715	K02051	AZC_2351	ABC-type nitrate/sulfonate/taurine/bicarbonate transport systems, periplasmic components
167868494	0.00004	COG1638		SMa0157	dicarboxylate-binding periplasmic protein
167866460	0.00004	COG0687	K02055	SCO5667	spermidine/putrescine-binding periplasmic protein
167719840	0.00004	COG0085	K03043	SAR11_1123	DNA-directed RNA polymerase beta subunit/140 kD subunit (split gene in Mjan, Mthe, Aful) : rpoB
167816636	0.00004	COG0459	K04077	Krad_0736	chaperonin GroEL (HSP60 family)
167701680*	0.00004	COG3740	K06904	BL0376	phage head maturation protease
167717794*	0.00003	COG0195	K02600	SAR11_0388	phage head maturation protease
167717838	0.00003	COG0443	K04043	SAR11_0368	molecular chaperone : dnaK
167834314	0.00003	COG0443		CMS_2806	molecular chaperone : dnaK
167717784	0.00003	COG1185	K00962	SAR11_0392	polyribonucleotide nucleotidyltransferase (polynucleotide phosphorylase)
167818278	0.00003	COG1022		Noca_3113	long-chain acyl-CoA synthetases (AMP-forming) : AMP-dependent synthetase and ligase
167719850	0.00002	COG0480	K02355	SAR11_1119	translation elongation and release factors (GTPases) : fusA
167816480	0.00002	COG0086	K03046	Krad_0681	DNA-directed RNA polymerase beta' subunit/160 kD subunit (split gene in archaea and Syn)
167866408	0.00001	COG1185	K00962	Lxx09030	polyribonucleotide nucleotidyltransferase (polynucleotide phosphorylase)
5 m - KEGG and NR annotated proteins					
Gene ID	NSA	NR ID	KO ID	KEGG/NR cus	KEGG/NR description
167873078	0.03710	BAF91544			major capsid protein [uncultured Myoviridae]
167771989 <i>i</i>	0.02658		BTH_I0914		hypothetical protein
167723550 <i>j</i>	0.01559	YP_001648266	OsV5_190f		hypothetical protein [Ostreococcus virus OsV5]
167927818 <i>j</i>	0.01345	YP_001648266	OsV5_190f		hypothetical protein [Ostreococcus virus OsV5]
167933090	0.01298	YP_002590925			putative porin [Candidatus Pelagibacter sp. HTCC7211]
167711088	0.01230		K09969	PBPRA2185	putative amino acid ABC transporter, periplasmic amino acid-binding protein
167691398	0.01175			HM1_2880	phage major capsid protein, hK97 family
167922719 <i>k</i>	0.01174			Neut_1469	phage major capsid protein, HK97 family protein

Continued on next page

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167775105 <i>l</i>	0.01103	ABC95191			GP23-major capsid protein [Stenotrophomonas phage SMB14]
167687982 <i>l</i>	0.00968	NP_944113			gp23 major head protein [Aeromonas phage Aeh1]
167748599 <i>m</i>	0.00960			M6_Spy1138	phage prohead protease
167733772	0.00952			BBta_5785	putative phage major head protein
167925660	0.00923			SRU_2178	putative outer membrane protein, probably involved in nutrient binding
167796059 <i>n</i>	0.00853			APEC01_525	hypothetical protein
167883590	0.00820			PP_1567	phage major capsid protein, HK97 family
167666520 <i>o</i>	0.00818	BAF91544			major capsid protein [uncultured Myoviridae]
167664173* <i>p</i>	0.00713			GDI3673	hypothetical protein
167667150 <i>m</i>	0.00687			LGAS_1485	predicted phage phi-C31 GP36 major capsid-like protein
167771337	0.00605	ABW90951			gp23 major capsid protein [uncultured Myoviridae]
167884290 <i>l</i>	0.00573	BAF91544			major capsid protein [uncultured Myoviridae]
167760139	0.00561			CHU_2679	probable outer membrane lipoprotein P61
167816468 <i>q</i>	0.00522			DR_A0099	hypothetical protein
167700634 <i>j</i>	0.00499	YP_001648266		OsV5_190f	hypothetical protein [Ostreococcus virus OsV5]
167687792 <i>r</i>	0.00498			Asuc_1240	phage major capsid protein, HK97 family
167842580 <i>r</i>	0.00453			BAV1464	major capsid protein
167700776	0.00447			Bpro_3745	hypothetical protein
167729766	0.00433	ZP_01224596		GB2207_03424	hypothetical protein [marine gamma proteobacterium HTCC2207]
167934698	0.00431			Swit_4452	hypothetical protein
167884738	0.00409			BBta_5785	putative phage major head protein
167669610 <i>p</i>	0.00397			GDI3673	hypothetical protein
167861688	0.00359			BDI_2874	putative outer membrane protein, probably involved in nutrient binding
167910063	0.00347			GDI3673	hypothetical protein
167893743* <i>s</i>	0.00338	YP_001648158		OsV5_081f	hypothetical protein [Ostreococcus virus OsV5]
167888926 <i>p</i>	0.00326			GDI3673	hypothetical protein
167753643*	0.00324			Daci_1946	putative phage major head protein
167742624 <i>p</i>	0.00317			GDI3673	hypothetical protein
167908539 <i>r</i>	0.00304			BAV1464	major capsid protein
167675286*	0.00284			CKO_01864	hypothetical protein
167900893 <i>n</i>	0.00278			APEC01_525	hypothetical protein
167778265 <i>p</i>	0.00275			GDI3673	hypothetical protein
167786471	0.00267			mlr8524	phage major capsid protein, GP36
167735768	0.00265			FRAAL2681	hypothetical protein

Continued on next page

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167773951t	0.00253			LGAS_1485	predicted phage phi-C31 GP36 major capsid-like protein
167841586j	0.00250	A7U6E7		Pmen_3970	putative major capsid protein [Chrysochromulina ericina virus]
167919545	0.00245				phage major capsid protein, HK97 family
167781901u	0.00236	YP_214367			T4-like major capsid protein [Prochlorococcus phage P-SSM2]
167923659p	0.00235			GDI3673	hypothetical protein
167852301v	0.00230			MAB_1788	bacteriophage protein
167659301	0.00224			LGAS_1485	predicted phage phi-C31 GP36 major capsid-like protein
167861686	0.00223	YP_002789013			TonB dependent/ligand-gated channel [Polaribacter sp. MED152]
167712528v	0.00215			MAB_1788	bacterial nucleoid DNA-binding protein
167678920*w	0.00209	YP_001498525		AR158_C444L	hypothetical protein [Paramecium bursaria Chlorella virus AR158]
167849540j	0.00201	YP_001648266		OsV5_190f	hypothetical protein [Ostreococcus virus OsV5]
167781903u	0.00201	YP_214367			T4-like major capsid protein [Prochlorococcus phage P-SSM2]
167863158j	0.00200	YP_001648266		OsV5_190f	hypothetical protein [Ostreococcus virus OsV5]
167663967*	0.00190			Swit_4461	hypothetical protein
167687108u	0.00182	YP_214367			T4-like major capsid protein [Prochlorococcus phage P-SSM2]
167692622i	0.00176			SG1188	hypothetical protein
167733858	0.00170	ZP_01017474			major capsid protein, HK97 family protein [Parvularcula bermudensis HTCC2503]
167852851p	0.00167			GDI3673	hypothetical protein
167864542k	0.00166			Neut_1469	phage major capsid protein, HK97 family protein
167803157	0.00165	ZP_01688540			lipoprotein, putative [Microscilla marina ATCC 23134]
167682644j	0.00153	A7U6E7			putative major capsid protein [Chrysochromulina ericina virus]
167733004*j	0.00150	A7U6F0			putative major capsid protein [Phaeocystis pouchetii virus]
167765429	0.00148			CHU_2610	gliding motility-related protein; possible GldN and/or GldO
167878228t	0.00145			LGAS_1485	predicted phage phi-C31 GP36 major capsid-like protein
167775103	0.00145	YP_214669			gp23 [Prochlorococcus phage P-SSM4]
167702908	0.00143			mma_2202	hypothetical protein
167869946	0.00135	YP_195142			major capsid protein gp23 [Synechococcus phage S-PM2]
167834518	0.00128			Haur_0657	hypothetical protein
167807747	0.00122			Saro_0657	hypothetical protein
167816420	0.00121			APECO1_525	hypothetical protein
167809283k	0.00119			Neut_1469	phage major capsid protein, HK97 family protein
167868514	0.00119			SAR11_1290	TRAP-type bacterial extracellular solute-binding protein
167750765	0.00118			Smed_1334	phage major capsid protein, HK97 family

Continued on next page

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167925457p	0.00115			GDI3673	hypothetical protein
167782759	0.00113			Oter_1957	band 7 protein
167871794j	0.00113	YP_001648266		OsV5_190f	hypothetical protein [Ostreococcus virus OsV5]
167756019	0.00112			CHU_0172	gldL; gliding motility-related protein
167670926x	0.00112			BBta_5785	putative phage major head protein
167821362z	0.00112			APECO1_525	hypothetical protein
167690910j	0.00111	YP_001648266		OsV5_190f	hypothetical protein [Ostreococcus virus OsV5]
167685332j	0.00110	YP_001648266		OsV5_190f	hypothetical protein [Ostreococcus virus OsV5]
167700460*	0.00110	YP_001648158		OsV5_081f	hypothetical protein [Ostreococcus virus OsV5]
167685474*aa	0.00104	YP_001648182		OsV5_105r	hypothetical protein [Ostreococcus virus OsV5]
167734326bb	0.00103			amb4267	hypothetical protein
16775653q	0.00101			Haur_0657	hypothetical protein
167733302p	0.00101			GDI3673	hypothetical protein
167663981p	0.00097			GDI3673	hypothetical protein
167763843	0.00096			Saro_0657	hypothetical protein
167768193z	0.00096			CKO_01864	hypothetical protein
167719228i	0.00095			SG1188	hypothetical protein
167844676	0.00091	ZP_03643684		BACCOPRO_02057	hypothetical protein [Bacteroides coprophilus DSM 18228]
167881504cc	0.00091			BSU26140	yqbE; hypothetical protein
167852559oo	0.00090			HSM_0907	hypothetical protein
167804465*	0.00088	ZP_03724502		ObacDRAFT_9001	hypothetical protein [Opitutaceae bacterium TAV2]
167794165p	0.00087			GDI3673	hypothetical protein
167734676	0.00085			Amet_4028	phage major capsid protein, HK97 family
167764813u	0.00084	YP_214367			T4-like major capsid protein [Prochlorococcus phage P-SSM2]
167781783p	0.00079			GDI3673	hypothetical protein
167759955	0.00077			Dgeo_0628	hypothetical protein
167733674	0.00076			Swit_4452	hypothetical protein
167878828	0.00075	K02027		SAV1394	ABC transporter solute-binding protein
167740142	0.00075	YP_002705257			gp34 [Stenotrophomonas sp. SKA14]
167834088	0.00074			Haur_0657	hypothetical protein
167821604j	0.00072	YP_001648266		OsV5_190f	hypothetical protein [Ostreococcus virus OsV5]
167798697p	0.00070			GDI3673	hypothetical protein
167823322s	0.00070	YP_001648158		OsV5_081f	hypothetical protein [Ostreococcus virus OsV5]
167718758j	0.00070	A7U6F0			putative major capsid protein [Phaeocystis pouchetii virus]

Continued on next page

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167713806	0.00069			SG1188	hypothetical protein
167778269 ^{dd}	0.00068	YP_002276820		Gdia_2460	hypothetical protein [Gluconacetobacter diazotrophicus PA1 5]
167879936 ^s	0.00067	YP_001648158		OsV5_081f	hypothetical protein [Ostreococcus virus OsV5]
167701850	0.00064			M446_5960	hypothetical protein
167934792 ^p	0.00064			GDI3673	hypothetical protein
167874674	0.00063			GFO_0492	conserved hypothetical protein, secreted-possibly porin
167821292	0.00062			Oant_1504	peptidase U35 phage prohead HK97
167867556	0.00062			Rru_A2587	hypothetical protein
167901481	0.00061			Cthe_1719	phage major capsid protein, HK97 family
167824444	0.00059			Smed_5134	TRAP dicarboxylate transporter-DctP subunit
167696166*	0.00059			BTH_I0915	hypothetical protein
167936648	0.00056	EEI06235		XcelDRAFT_1815	hypothetical protein [Xylanimonas cellulosilytica DSM 15894]
167910361 ^j	0.00054	A7U6F0			putative major capsid protein [Phaeocystis pouchetii virus]
167675492 ^{*p}	0.00054			GDI3673	hypothetical protein
167801933 ^{aa}	0.00052	YP_001648182		OsV5_105r	hypothetical protein [Ostreococcus virus OsV5]
167725772 ^{cc}	0.00051			BSU26140	yqbE; hypothetical protein
167820670	0.00050			gll0198	similar to bacteriorhodopsin
167832972 ^t	0.00050			LGAS_1485	predicted phage phi-C31 GP36 major capsid-like protein
167867536	0.00049			TM1040_0812	hypothetical protein
167783747	0.00048			Glov_2914	cell surface receptor IPT/TIG domain protein
167893945	0.00048			Oter_3420	hypothetical protein
167740708	0.00047		K03286	Pnap_1319	OmpA/MotB domain protein; OmpA-OmpF porin, OOP family
167772783	0.00047			RCIX1696	hypothetical protein
167734614 ^p	0.00047			GDI3673	hypothetical protein
167776587*	0.00046	YP_001648249		OsV5_172f	hypothetical protein [Ostreococcus virus OsV5]
167911245 ^s	0.00046	YP_001648315		OsV5_239r	hypothetical protein [Ostreococcus virus OsV5]
167867748 ^x	0.00044			Daci_1946	putative phage major head protein
167732694	0.00044			NMC0858	putative phage-related protein
167922873	0.00042			CHU_3230	hypothetical protein
167734178 ^p	0.00041			GDI3673	hypothetical protein
167873260	0.00041			Sare_3763	hypothetical protein
167685638 ^{*w}	0.00041	YP_001498525		AR158_C444L	hypothetical protein [Paramecium bursaria Chlorella virus AR158]
167901149	0.00040		K01358	azo1870	endopeptidase Clp; K01358 ATP-dependent Clp protease, protease subunit
167853099 ⁱ	0.00040			SG1188	hypothetical protein

Continued on next page

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167761349x	0.00037			Daci_1946	putative phage major head protein
167776241	0.00037			mll0455	hypothetical protein
167824154	0.00036			SACE_4894	hydrolase, alpha/beta fold family
167843578s	0.00035	YP_001648158		OsV5_081f	hypothetical protein [Ostreococcus virus OsV5]
167682808*	0.00035	YP_001648239		OsV5_162f	hypothetical protein [Ostreococcus virus OsV5]
167703228	0.00034			Dshi_0412	beta-Ig-H3/fasciclin
167918033p	0.00034			GDI3673	hypothetical protein
167891224p	0.00033			GDI3673	hypothetical protein
167867622	0.00033			Oant_1504	peptidase U35 phage prohead HK97
167732430ff	0.00031	ZP_02092868		FAEPRAM212_03171	hypothetical protein [Faecalibacterium prausnitzii M21/2]
167684500s	0.00031	YP_001648158		OsV5_081f	hypothetical protein [Ostreococcus virus OsV5]
167824164*	0.00030	YP_001648240		OsV5_163f	hypothetical protein [Ostreococcus virus OsV5]
167765431	0.00030			CHU_0173	gldM; gliding motility-related protein
167854137	0.00030	ZP_00743477		RBTH_08297	hypothetical protein [Bacillus thuringiensis serovar israelensis ATCC 35646]
167730288k	0.00028			Neut_1469	phage major capsid protein, HK97 family protein
167685472*gg	0.00027	YP_001648184		OsV5_107r	hypothetical protein [Ostreococcus virus OsV5]
167778267p	0.00027			GDI3673	hypothetical protein
167919557	0.00027			PputW619_3936	hypothetical protein
167782645	0.00026			BBta_5785	putative phage major head protein
167908551	0.00026			CC_2781	hypothetical protein
16786723p	0.00026			GDI3673	hypothetical protein
167896531	0.00026			BAV1464	major capsid protein
167821374	0.00025	YP_001919460		Mpop_5468	hypothetical protein [Methylobacterium populi BJ001]
167833472	0.00024			GDI3673	hypothetical protein
167733210	0.00024			Pmen_3970	phage major capsid protein, HK97 family
167713652*	0.00023			mlr8533	hypothetical protein
167935700p	0.00023			GDI3673	hypothetical protein
167872214hh	0.00023	K06907		Sfum_3815	phage tail sheath protein
167881636	0.00023			Rspf17025_0103	hypothetical protein
167791200p	0.00023			GDI3673	hypothetical protein
167922981p	0.00022			GDI3673	hypothetical protein
167824604	0.00022			Rspf17029_3578	uncharacterized protein UPF0065
167933608ff	0.00022	ZP_02092868		FAEPRAM212_03171	hypothetical protein [Faecalibacterium prausnitzii M21/2]
167817058	0.00022		K00518	Sare_4077	superoxide dismutase

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167823358*gg	0.00021	YP_001648184		OsV5_107r	hypothetical protein [Ostreococcus virus OsV5]
167892855	0.00021	YP_001648184		OsV5_107r	hypothetical protein [Ostreococcus virus OsV5]
167840790	0.00021			Rspf17025_0437	hypothetical protein
167712150*s	0.00021	YP_001648158		OsV5_081f	hypothetical protein [Ostreococcus virus OsV5]
167833160bb	0.00021			amb4267	hypothetical protein
167892985j	0.00021	YP_001648266		OsV5_190f	hypothetical protein [Ostreococcus virus OsV5]
167696294oo	0.00020			HS_1377	hypothetical protein
167759041	0.00019			PP_3877	hypothetical protein
167766087s	0.00019	YP_001648153		OsV5_076f	hypothetical protein [Ostreococcus virus OsV5]
167919775	0.00019	YP_001294637		ORF044	hypothetical protein [Pseudomonas phage PA11]
167804453*	0.00017	ZP_03724505		ObacDRAFT_9004	hypothetical protein [Opitutaceae bacterium TAV2]
167833104	0.00017			Bcep1808_1173	hypothetical protein
167721370*	0.00016	YP_001648301		OsV5_225r	hypothetical protein [Ostreococcus virus OsV5]
167826943*	0.00016			Sare_3763	hypothetical protein
167865492	0.00015		K02027	Pput_3473	extracellular solute-binding protein, family 1; multiple sugar transport system substrate-binding protein
167685780*j	0.00015	YP_001648266		OsV5_190f	hypothetical protein [Ostreococcus virus OsV5]
167910713	0.00014			Bd1641	hypothetical protein
167910061p	0.00014			GDI3673	hypothetical protein
167833358s	0.00013	YP_001648158		OsV5_081f	hypothetical protein [Ostreococcus virus OsV5]
167692314*w	0.00012	YP_001498525		AR158_C444L	hypothetical protein [Paramecium bursaria Chlorella virus AR158]
167925393	0.00012			Oter_3421	hypothetical protein
167687436	0.00011		K01999	azo3443	conserved hypothetical ABC-type branched-chain amino acid transport systems, periplasmic component
167719658hh	0.00011		K06907	Dde_1889	hypothetical protein
167668360s	0.00011	YP_001648158		OsV5_081f	hypothetical protein [Ostreococcus virus OsV5]
167735772	0.00010			FRAAL2683	hypothetical protein; putative mycobacteriophage protein (GP15) similarity
167702102	0.00010			Daci_1946	putative phage major head protein
167688622p	0.00009			GDI3673	hypothetical protein
167782867cc	0.00009			BSU26140	yqbE; hypothetical protein
167867386	0.00008			TM1040_1299	peptidase U35, phage prohead HK97
167789595	0.00008			APECO1_4044	hypothetical protein
167828425*w	0.00007	YP_001498525		AR158_C444L	hypothetical protein [Paramecium bursaria Chlorella virus AR158]

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167865490	0.00007		K02027	Rmet_2229	extracellular solute-binding protein, family 1; multiple sugar transport system substrate-binding protein
167840812	0.00006	ZP_01959135		BACCAC_00731	hypothetical protein [Bacteroides caccae ATCC 43185]
167706428	0.00005	YP_001648190		OsV5_113r	hypothetical protein [Ostreococcus virus OsV5]
167867920 <i>p</i>	0.00005			GDI3673	hypothetical protein
167842648* <i>s</i>	0.00005	YP_001648315		OsV5_239r	hypothetical protein [Ostreococcus virus OsV5]
167869096 <i>w</i>	0.00005	YP_001498525		AR158_C444L	hypothetical protein [Paramecium bursaria Chlorella virus AR158]
167859444	0.00005	YP_001648151		OsV5_074f	hypothetical protein [Ostreococcus virus OsV5]
167871600	0.00005	YP_001648152		OsV5_075f	hypothetical protein [Ostreococcus virus OsV5]
167669608 <i>p</i>	0.00004			GDI3673	hypothetical protein
167678686 <i>p</i>	0.00004			GDI3673	hypothetical protein
167752119	0.00004	YP_001648152		OsV5_075f	hypothetical protein [Ostreococcus virus OsV5]
167825992 <i>j</i>	0.00003	A7U6E7			putative major capsid protein [Chrysochromulina ericina virus
167818634	0.00003			PputGB1_1751	hypothetical protein
167671778*	0.00003	YP_001648124		OsV5_047f	hypothetical protein [Ostreococcus virus OsV5]
167871626 <i>s</i>	0.00002	YP_001648158		OsV5_081f	hypothetical protein [Ostreococcus virus OsV5]
167753841	0.00002		K06907	Sfum_3815	phage tail sheath protein
167724632*	0.00002	YP_001648232		OsV5_155f	hypothetical protein [Ostreococcus virus OsV5]
167690816	0.00002	YP_001648190		OsV5_113r	hypothetical protein [Ostreococcus virus OsV5]
167742884*	0.00001			Dvul_0646	hypothetical protein
167875342 <i>j</i>	0.00001	YP_001648145		OsV5_068f	hypothetical protein [Ostreococcus virus OsV5]
5 m - Proteins with no annotation					
167699580*	0.01263				
167736790 <i>ii</i>	0.01043				
167796769	0.00914				
167722626 <i>jj</i>	0.00789				
167703824 <i>pp</i>	0.00714				
167753801	0.00626				
167854251	0.00577				
167744898* <i>o</i>	0.00546				
167664175* <i>nn</i>	0.00514				
167829145 <i>pp</i>	0.00433				
167779175 <i>pp</i>	0.00423				
167881060 <i>pp</i>	0.00419				

Continued on next page

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167887022 <i>pp</i>	0.00390				
167836216 <i>jj</i>	0.00369				
167688044 <i>p</i>	0.00353				
167697984 <i>pp</i>	0.00321				
167855765 <i>jj</i>	0.00318				
167764897 <i>jj</i>	0.00297				
167718436	0.00240				
167771817	0.00226				
167699330*	0.00220				
167891152 <i>pp</i>	0.00207				
167844558 <i>pp</i>	0.00197				
167801097 <i>pp</i>	0.00197				
167891908	0.00196				
167820168 <i>ii</i>	0.00192				
167688624	0.00182				
167746546 <i>jj</i>	0.00176				
167682238 <i>p</i>	0.00175				
167722606	0.00164				
167883488 <i>pp</i>	0.00157				
167839862 <i>mm</i>	0.00139				
167820406*	0.00139				
167858104	0.00138				
167806741 <i>jj</i>	0.00138				
167678192	0.00133				
167706644	0.00133				
167787801* <i>ll</i>	0.00124				
167781039	0.00124				
167936638	0.00116				
167733354	0.00116				
167918031	0.00112				
167790652	0.00110				
167734428	0.00102				
167925455	0.00102				
167928078	0.00100				

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167682970o	0.00098				
167701282	0.00091				
167867140o	0.00087				
167809975jj	0.00087				
167750727	0.00082				
167883564ll	0.00082				
167789467pp	0.00080				
167669606	0.00078				
167733300	0.00076				
167750389jj	0.00072				
167852849	0.00072				
167827017	0.00070				
167691436	0.00068				
167816466*	0.00067				
167678688	0.00063				
167796679	0.00062				
167761163	0.00059				
167916021pp	0.00059				
167867918	0.00058				
167853885	0.00058				
167757667	0.00053				
167922983	0.00052				
167923663nn	0.00051				
167922109	0.00051				
167661777	0.00051				
167936684	0.00050				
167867228	0.00050				
167791202	0.00049				
167819274	0.00046				
167765833*	0.00045				
167793451*	0.00045				
167732910	0.00043				
167890226	0.00043				
167718438	0.00042				

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167688708	0.00041				
167699600 <i>pp</i>	0.00041				
167746630	0.00040				
167821290	0.00039				
167916161*	0.00038				
167700778	0.00037				
167701632 <i>kk</i>	0.00037				
167675494*	0.00036				
167711820	0.00034				
167663983	0.00033				
167689444*	0.00032				
167933464 <i>nn</i>	0.00032				
167891222	0.00032				
167852557	0.00031				
167843828	0.00031				
167843020	0.00031				
167677672	0.00030				
167776503*	0.00028				
167804815	0.00027				
167713808	0.00027				
167702310 <i>dd</i>	0.00026				
167913463o	0.00025				
167881302	0.00025				
167907624 <i>mm</i> 0.00025					
167753609*	0.00024				
167829571	0.00024				
167921665	0.00024				
167920645	0.00023				
167714058	0.00022				
167677546*	0.00021				
167913465	0.00020				
167697624*	0.00020				
167912083 <i>jj</i>	0.00019				
167766043*	0.00018				

Continued on next page

Table 6.2 – *Continued from previous page*

Gene ID	NSA	COG ID	KO ID	KEGG locus	COG description:KEGG description
167678558*	0.00018				
167733556	0.00017				
167663383*	0.00016				
167905220*	0.00015				
167891594	0.00015				
167883594 <i>ll</i>	0.00015				
167879460	0.00014				
167919777	0.00014				
167884588 <i>o</i>	0.00014				
167822810	0.00013				
167713494	0.00012				
167841896*	0.00010				
167804467*	0.00010				
167925043 <i>kk</i>	0.00010				
167858106	0.00009				
167788223*	0.00009				
167878206	0.00008				
167764895 <i>jj</i>	0.00008				
167767179	0.00008				
167858640	0.00008				
167683530*	0.00007				
167918035	0.00006				
167766125*	0.00006				
167752051*	0.00005				
167890228	0.00004				
167685654*	0.00003				
167719670*	0.00002				
167879450	0.00002				